DOCUMENT RESUME

ED 353 294	TM 019 284
TITLE	Assessing Student Achievement in the States: Background Studies. Studies for the Evaluation of the NAEP Trial State Assessment Commissioned for the National Academy of Education Panel Report on the 1990 Trial.
INSTITUTION	National Academy of Education, Stanford, Calif.
PUB DATE	92
NOTE	439p.
AVAILABLE FROM	National Academy of Education, Stanford University, School of Education, CERAS-507, Stanford, CA 94305-3084 ($10 prepaid).
PUB TYPE	Collected Works - General (020)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Academic Achievement; Achievement Tests; *Educational Assessment; Eligibility; Grade 8; Inferences; Junior High Schools; Junior High School Students; Mathematics Tests; *National Surveys; Program Evaluation; Program Implementation; Public Schools; Research Reports; State Programs; *Student Evaluation; Testing Programs; Test Results
IDENTIFIERS	*Trial State Assessment (NAEP)

ABSTRACT

This document presents 10 papers concerning the 1990 Trial State Assessment, which was restricted to the eighth-grade public school population in 37 states, the District of Columbia, and 2 U.S. territories. The following are the papers: (1) "A Critique of Sampling in the 1990 Trial State Assessment" (Bruce D. Spencer); (2) "Eligibility/Exclusion Issues in the 1990 Trial State Assessment" (Bruce D. Spencer); (3) "Evaluation of the Implementation of the 1990 Trial State Mathematics Assessment" (Donald H. McLaughlin, Frances B. Stancavage, Jay G. Chambers, Elizabeth Hartka, and Kadriye Ercikan); (4) "The Content and Curricular Validity of the 1990 NAEP Mathematics Items: A Retrospective Analysis" (Edward A. Silver, Patricia A. Kenney, and Leslie Salmon-Cox); (5) "The Relative Standing of States in the 1990 Trial State Assessment: The Influence of Choice of Content, Statistics, and Subpopulation Breakdowns" (Robert L. Linn, Lorrie Shepard, and Elizabeth Hartka); (6) "A Study of the Impact of Reporting the Results of the 1990 Trial State Assessment: First Report" (Frances B. Stancavage, Edward Roeber, and George W. Bohrnstedt); (7) "General Issues in Reporting the Results of the NAEP Trial State Assessment Results" (Robert M. Jaeger); (8) "The Case for District- and School-Level Results from NAEP" (Ramsey Selden); (9) "Cautions on the Future of NAEP: Arguments Against Using NAEP Tests and Data Reporting below the State Level" (Walter Haney and George F. Madaus); and (10) "Reasonable Inferences for the Trial State NAEP Given the Current Design: Inferences that Can and Cannot Be Made" (Edward H. Haertel). (RLC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

ED353294

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
L.S. SHULMAN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Assessing Student Achievement in the States: Background Studies

Studies for the Evaluation
of the NAEP Trial State Assessment
Commissioned for the National Academy of Education Panel
Report on the 1990 Trial

Chairmen
Robert Glaser, University of Pittsburgh
Robert Linn, University of Colorado

Project Director
George Bohrnstedt, American Institutes for Research

TM019284

BEST COPY AVAILABLE

Assessing Student Achievement in the States: Background Studies

Copyright © 1992 by the National Academy of Education
All rights reserved.
Printed in the United States of America
Single copy (prepaid): $10
Assessing Student Achievement in the States—The First Report of the Panel: $20
National Academy of Education
Stanford University
School of Education CERAS-507
Stanford CA. 94305-3084
(415) 725-1003

Assessing Student Achievement in the States: *Background Studies*

Studies for the Evaluation
of the NAEP Trial State Assessment
Commissioned for the National Academy of Education Panel
Report on the 1990 Trial

Contributors

George Bohrnstedt
Jay Chambers
Kadriye Ercikan
Edward Haertel
Walter Haney
Elizabeth Hartka
Richard Jaeger
Patricia Ann Kenney
Robert Linn
Donald McLaughlin
George Madaus
Edward Roeber
Leslie Salmon-Cox
Ramsay Selden
Edward Silver
Bruce Spencer
Lorrie Shepard
Frances Stancavage

Table of Contents

A Critique of Sampling in the 1990 Trial State Assessment by Bruce D. Spencer........................1

Introduction 3
Sampling of Schools 4
Sampling of Students 6
Weighting and Sampling Errors..7
Summary and Recommendations 17
References 18

Eligibility/Exclusion Issues in the 1990 Trial State Assessment by Bruce D. Spencer................19

Introduction 21
Who Was Supposed to Be Assessed?...22
Who Should Be Assessed? 46
Summary and Recommendations 47
References 49

Evaluation of the Implementation of the 1990 Trial State Mathematics Assessment
by Donald H. McLaughlin, Frances By Stancavage, Jay G. Chambers,
Elizabeth Hartka, and Kadriye Ercikan..51

Executive Summary 53
Introduction 59
Validity Estimation...63
Feasibility Evaluation 118
Summary and Conclusions 134
Appendix A: Variance Estimation and Significance Testing....................141
Appendix B: Survey Instruments Used by AIR.......................................145

The Content and Curricular Validity of the 1990 NAEP Mathematics Items: A Retrospective Analysis by Edward A. Silver, Patricia Ann Kenney, and Leslie Salmon-Cox 157

Introduction 159
Analysis of the Consensual Process 161
Analysis of the Validity of the Framework and Items 166
Recommendations 185
Appendix A: CCSSO Committee Members and Consultants 193
Appendix B: People Interviewed for the Study Examining the Consensual Process 199
Appendix C: Letter to State-Level Mathematics Specialists/Consultants 203
Appendix D: Telephone Interview Protocol 207
Appendix E: Biographical Sketches of Expert Panel (EP) Members 211
Appendix F: Themes from the NCTM *Standards* for Grades 5-8 215

The Relative Standing of States in the 1990 Trial State Assessment: The Influence of Choice of Content, Statistics, and Subpopulation Breakdowns by Robert L. Linn, Lorrie Shepard, and Elizabeth Hartka 219

Introduction 221
Method 225
Results 228
Conclusions and Recommendations 256
References 257

A Study of the Impact of Reporting the Results of the 1990 Trial State Assessment: First Report by Frances B. Stancavage, Edward Roeber, and George W. Bohrnstedt 259

Purpose of the Study 261
Methodology 262
Research Findings 269
Summary and Conclusions 283

General Issues in Reporting the Results of the NAEP Trial State Assessment Results by Richard M. Jaeger 285

Executive Summary 287
Background 289
Requirements of the 1985 *Standards for Educational and Psychological Testing* 292
Reporting to the Public 295
A Reprise, a Recommendation, and an Observation 332
Reporting to Policymakers 335
Conclusions and Recommendations 339
References 341

The Case for District- and School-Level Results from NAEP by Ramsey Selden 345

Background 347
Arguments for School- and District-Level NAEP 348
Liabilities in a School- and District-Level NAEP 357
Overall Assessment of Impact 360
References 362

Cautions on the Future of NAEP: Arguments Against Using NAEP Tests and Data Reporting Below the State Level by Walter Haney and George F. Madaus 363

I. Introduction 365
II. Proposals to Remove the Prohibition on Reporting NAEP below the State Level 368
III. Reasons behind the Original Reporting Prohibitions on NAEP 381
IV. Practical Problems 389
V. The Problem of High-Stakes Testing 393
VI. First Things First in Educational Policymaking 398
References 402

Reasonable Inferences for the Trial State NAEP Given the Current Design: Inferences That Can and Cannot Be Made by Edward H. Haertel 409

Abstract 411
Introduction 413
Policy Context 414
Structural and Technical Characteristics of the TSA Data 418
Warranted and Unwarranted Substantive Inferences from the TSAs 433
Summary 442
Recommendations 444
References 446

A Critique of Sampling in the 1990 Trial State Assessment

Bruce D. Spencer
Methodology Research Center, NORC
Northwestern University

November 1991

Acknowledgements

The author is grateful to Jackie Belanger, George Bohrnstedt, Bob Groves, Lyle Jones, Thomas Kerins, Bob Linn, Keith Rust, and Lorrie Shepard for their generous comments and assistance.

Introduction

The 1990 Trial State Assessment (TSA) was restricted to the eighth grade public school population in 37 states, Washington, DC, and 2 U.S. territories. The survey excluded certain large groups of students, some by design (e.g.,private school students and students with limited English proficiency or individualized education plans) and some by chance (nonparticipants, movers who "fell through the cracks" of the survey procedure, and absentees). This report discusses the design and implementation of the sample, including issues of nonparticipation by students who were intended to participate in the TSA. Another report (Spencer 1991) discusses issues related to the exclusions by design.

The study was charged with examining the following:

1. specification of target of estimation (if 100 percent sampling were used),
2. adequacy of sampling frame, both at institution level and at person level,
3. implementation of sample design; uses of stratification; variability of designs among states,
4. poststratification,
5. issues in merging national and state samples, particularly weighting,
6. efficiency of sampling design and weighting,
7. nonresponse, at state, institution, person, and item levels,
8. adequacy of methods for estimating standard errors and other measures of uncertainty,
9. adequacy of methods for adjusting for missing data, both weighting and imputation,
10. accuracy of certain statistics
11. presentation of statistical results (conformity with ASA guidelines),
12. nonrandom variations, including eligibility criteria and others as appropriate,
13. effects of exclusion of private schools.

Topics 1-10 are discussed in this report. Topic 11 was not examined because there were too few presentations to assess. The issues of exclusion, 12-13, are discussed in a separate report on eligibility and exclusion issues (Spencer 1991).

Sampling of Schools

The sampling design called for sampling about 100 schools per state; if a state had fewer than 100 schools then all were selected (unless, as in Wyoming, there were very large numbers of very small schools, in which case the small schools were sampled). In order to ensure that, within each state, students would have approximately equal chances of being selected in the sample, the schools were selected with probabilities proportional to their estimated eighth grade enrollment. Special procedures were used for small schools (thought to have fewer than 20 students).

The list of schools was based on the National Center for Education Statistics (NCES) Common Core of Data for 1988 for each state. The states' samples of schools were selected by systematic stratified sampling. The strata were based on urbanicity (central city, suburban, other), percent of students who were black or Hispanic, median income, and eighth grade enrollment size. This is standard stratification practice and is appropriate for broad-based surveys. Different sampling designs (i.e., different stratification and different methods of handling small schools) were used in different states. It is reasonable to expect some efficiency gains if the stratification could be based on prior school-level test scores, such as from statewide exams.

Some school districts had new schools that were not listed in the Common Core of Data. To allow for selection of these schools into the sample, any district that was selected into the sample (because one of its eighth grade schools was) was asked to provide information about new schools that were not on the sampling frame. Then these schools were subsampled so their selection probabilities too were proportional to their eighth grade enrollment sizes.

The sampling in all except the smaller states was performed so as to minimize the overlap with schools selected either for 1990 NAEP or 1990 NELS:88. The contractor properly calculated the selection probabilities for the selected schools with this in mind.

In all, the school sample design was the standard kind of sample design used in high quality major national education surveys. It seems to have been executed with care and due attention to details. *In future assessments, attention should be given to stratifying schools within a state by test scores on statewide exams.* This stratification approach should be empirically evaluated in at least a few states.

The overall participation rate for schools was about 94 percent, but it was less than 80 percent in two states, Illinois (78 percent) and Oklahoma (76 percent) (Rust 1991, Table 3). For each school whose nonparticipation was ascertained by November 1989, a single substitute school was designated. The substitute was chosen to resemble the original school with respect to the stratification criteria; if possible, the substitute was chosen from the same district. In all, about 188 substitutes were contacted and 99 of them participated. Even after substitution, some selected schools neither participated nor were "represented" by a substitute. Substitution increased the overall participation rate to 97 percent, although the rates were as low as 86 percent in New York. Weighting adjustments for school nonresponse were implemented.

A simple analysis was conducted to consider the effect of school nonparticipation in the TSA in Illinois. State assessment mean scores from the 1990 Illinois Goal Assessment Program eighth grade mathematics data were obtained for all schools eligible for the 1990 TSA in Illinois and were transformed to standardized scores[1]; thus the state mean score was 0. The mean[2] score for the (105) originally selected schools was -0.077. After substitution for 19 of the 23 nonparticipating schools, the mean changed to -0.050. The effect of school nonparticipation (as indicated by IGAP mean scores, was minor.[3] *In future assessments, evaluations of nonparticipation (nonresponse) bias should be performed at the school and student level.*

---

[1] The standardization was based on a simplifying assumption that there was no within-school variation in scores. This assumption was necessitated by the lack of data on student-level score variation.

[2] All sample means here are weighted by the reciprocal of the school selection probability.

[3] In fact, the nonparticipation bias just happened to reduce some of the effect of the sampling error, although this would not occur in general.

More elaborate and informative analyses than the one reported here could and should be done within the assessment program.

Sampling of Students

In each selected school a systematic sample of students was drawn from a list provided by the school. The sample size was usually 30 (or smaller if the school was smaller, although sometimes it was 60 or 100). The list was provided by December 1989, but the assessment occurred months later. This time lag meant that some students who were originally selected would have moved to other schools prior to assessment. To allow these students a proper chance to be tested, each school was to provide a supplementary list containing the names of all students who had moved into the school since the previous list was produced. Students were then sampled from the supplementary list at the same rate as the original sampling rate, and those selected were added to the sample.

The supplementary sampling procedure seems to have accounted for only 60 percent of the students it was supposed to cover. According to the contractor's report (Westat 1990), 4.9 percent of the students withdrew from the sample but the supplemental sampling procedure added only 2.9 percent to the sample. The other 2 percent (=4.9 percent - 2.9 percent) of the sample (40 percent of the withdrawals) seem to have been missed by the survey. *The undercoverage of transfer students should receive attention for future assessments.*

The participation rates of the selected students generally was good, ranging from 92 percent to 96 percent except in two states, where it was 88 percent and 80 percent.[4] The latter state required a signed statement of parental consent for each student who would participate, which explains the low participation rate. In most states, absentees comprised the major portion of the nonparticipating students in the sampled (and participating) schools. Make-up sessions were offered but they only captured about 1/5 of the students who did not participate in the initial testing. It is well known that participating and nonparticipating groups tend to differ for many kinds of characteristics. For example, there is evidence from other surveys that nonparticipating students tend to be lower scorers on tests (when they take the tests). It would have been interesting to compare the performance of absentee students who took the make-up test with the performance of other tested students. However, the computerized data files do not indicate whether a tested student took the make-up test or not and so this analysis was not feasible. *An analysis of the performance of absentee students taking make-up tests should be conducted in future assessments.* Differences in the demographic composition of participant groups and nonparticipant groups of students are also commonly observed in national educational surveys. For example, participation rates for whites typically are higher than participation rates for blacks. Student participation rates were calculated for three of the areas with the lowest rates, Oklahoma, Washington, DC, and New York, and this pattern was observed. Weighting adjustments for student nonresponse were implemented (see below). There is no reason to believe that ETS fully adjusted for the differences in the (potential) test scores of the participating and nonparticipating students, but neither is there evidence that the remaining systematic error from student nonparticipation is large. *The magnitude of the effect of student nonresponse warrants further investigation in future assessments.* Such investigations could use externally available test data for the nonresponding students, much in the way that test data for nonresponding schools in Illinois was used in the analysis described above.

---

[4] These rates represent the proportion of sampled students who participated in the assessment. The rates ignore students in schools that declined to participate.

Students who were classified as IEP or LEP *and* judged incapable of participating in the assessment were excluded from the assessment. This is not a sampling issue but a definition issue, since for better or worse the TSA statistics do not aim to reflect the performance of such students; (see Spencer 1991 for discussion).

Weighting and Sampling Errors

Weighting serves several purposes in complex surveys such as the TSA: it adjusts for unequal selection probabilities of students, it attempts to adjust for nonparticipation, it may attempt to increase precision (as in poststratification), and it can be used to implement resampling methods of estimating sampling errors (such as the jackknife and balanced repeated replication methods). In the TSA the adjustment for unequal selection probabilities is reasonably straightforward.

Weighting classes were used to adjust for school-level and student-level nonresponse. The basic assumption behind the use of a weighting class is that within a weighting class persons "decide" to participate or not at random; thus, within any weighting class (or cell) the mean score for participants is expected to be the same as what the mean score for nonparticipants would be if they had been assessed. The effectiveness of nonresponse weighting adjustments essentially depends on how well this assumption describes reality.[5] The weighting classes for the Trial State Assessment were based on stratification variables at the school level and demographic variables at the student level. The school-level weighting classes were formed for each state separately, which is appropriate, and were based on urbanicity and minority strata; if no minority strata were used in a state, then three equal-sized income classes (high, medium, and low) were used.[6]

Weighting classes for students were also based on characteristics only somewhat correlated with test scores. Students in a state were assigned to a weighting cell based on the weighting cell of the school they attended, whether they attended a monitored or unmonitored session, and whether they were less than 13 years old or at least 13.[7] Thus, the only student-level characteristic was an age dichotomy. Given well-documented differences in scores for different sexes and racial-ethnic groups, as well as inevitable differences in participation for those groups, the nonresponse adjustments cannot be expected to be very effective for removing bias due to nonparticipation by students.

The main reason that the weighting classes used for student-level nonresponse adjustments were so crude is that characteristics of nonparticipating students were not recorded. There is no reason to believe that the adjustments were highly effective in removing nonresponse bias. *In future assessments, attention should be*

---

[5] For empirical studies of the effectiveness of nonresponse adjustments, see Madow, Nisselson, and Olkin (1983).

[6] If all of the certainty selections participated, then the weighting cells were based only on noncertainty selections.

[7] Where necessary, weighting classes were combined to keep the cell sizes from becoming too small. Careful attention was given to the possibility that some of the student nonresponse adjustment factors might be too large and thereby cause overly large increases in sampling variance, but the analysis indicated that no special modifications were needed.

given to collecting data for nonparticipating students for the purposes of nonresponse adjustments. These characteristics should be chosen to be correlated both with propensity to participate and to test performance; examples include grade-point average, sex, race, and eligibility for subsidized lunch at school.[8] Attention should also be given to using outside test data for forming weighting cells for students. More effective nonresponse adjustments at the student level might well result. Care would need to be taken that schools do not corrupt the records of the outside test data, however.[9]

Special "comparison weights" were produced to increase the precision of estimates of differences between monitored and unmonitored sessions. These were developed by "raking" (i.e., adjusting sample estimates to conform to population marginal distributions) the usual student weights according to three sets of student classifications: race-ethnicity cross-classified by parents' education, age cross-classified by sex, and type of mathematics course taken cross-classified by self-perceived math proficiency.

In addition, a set of "replicate weights" was provided for each student for use in estimating standard errors by the jackknife method. The jackknife method is appropriate for use for estimating standard errors for the TSA, and the replicate weights are reasonably convenient to use. The standard errors are calculated under the assumption that the items on the test are fixed. If the items were viewed as representative of larger, possibly hypothetical pools of items, then the computations of sampling error would need to include an additional component, to reflect the sampling of items. Before such computations could be made, however, it would be necessary to specify how the "sampling" of items occurred. Another component of sampling error that is not reflected in the calculations but should be is error in the estimates of the item parameters in the IRT models. A major component of such parameter-estimation error may be errors in equating test forms (Johnson, Mislevy, and Zwick 1990).

---

[8] The nonresponse adjustment procedures could vary from state to state; consequently uniform availability of data is not an issue.

[9] Specifically, if the outside test scores for the respondents were deflated relative to the nonrespondents' scores, then the nonresponse adjustments would erroneously increase the estimate of the state's mean score.

The magnitude of the sampling error (ignoring item-sampling errors and parameter-estimation errors) for state-level estimates and rankings of states is reasonable. The average standard error of state mean scores (in standard units) is 0.024. Table 1 presents the standard errors for mean scores by state and subgroups within states. Table 2 reports 95 percent confidence intervals for the states' ranks (1 best and 40 worst) based on their mean scores; the average width of the confidence intervals is four, although the intervals for the states near the top and the states near the bottom seem to be narrowest. The graph following Table 2 shows how the intervals are widest for states whose observed ranks are in the middle 75 percent, (i.e., 5 - 35).

The magnitudes of sampling errors for proportions of students whose scores exceeded various proficiency levels are shown in Tables 3 - 5 for "basic," "proficient," and "advanced" levels, respectively. Rather than report the standard deviation, the tables report an estimate of the ratio of the standard error to the proportion being estimated. This ratio, known as the "coefficient of variation," expresses the sampling error in relative terms. For example, the first line in Table 5 shows that the proportion of students at the advanced level in state 1 is estimated to be 0.2 percent and the coefficient of variation of this estimate is 41.4 percent. Thus, the sampling error is about 40 percent of the estimate, or 0.08 percent. The coefficients of variation are quite high (often 0.3 or higher) for the proportion at the advanced level, and they are moderately high for the proportion at the proficient level. The acceptability of the level of sampling error must ultimately depend on the needs of the particular analysis into which the estimated proportions enter.

Table 1. **Sampling errors for state mean standard scores**

Standard errors for subgroups were estimated under the assumption that the design effect for subgroup mean scores would equal .9 times the design effect for the overall mean and that the proportion of tested students in a subgroup was equal to the proportion observed in the TSA. The factor of .9 is based on the assumption that effect of smaller average cluster sizes would slightly outweigh the increased intracluster correlations of subgroup scores.

			STANDARD ERRORS FOR SUBGROUPS					
	MEAN	SE	WHITE	BLACK	HISPAN	ASIAN	AM.IND.	OTHER
1	−0.322	0.028	0.034	0.048	0.123	0.288	0.162	1.413
2	−0.179	0.023	0.023	0.125	0.110	0.278	0.148	0.706
3	−0.244	0.027	0.032	0.048	0.121	0.253	0.253	—
4	−0.174	0.006	0.008	0.036	0.008	0.045	0.016	0.087
5	0.340	0.034	0.035	0.116	0.157	0.244	0.313	0.957
6	−0.419	0.032	0.042	0.050	0.133	0.313	0.286	1.726
7	0.009	0.033	0.036	0.095	0.136	0.233	0.104	1.610
8	−1.069	0.001	0.011	0.002	0.003	0.020	0.019	0.037
9	0.055	0.031	0.034	0.082	0.132	0.218	0.198	0.585
10	0.424	0.032	0.031	0.214	0.145	0.326	0.251	0.926
11	−0.754	0.003	0.012	0.026	0.007	0.004	0.032	—
12	−0.059	0.003	0.003	0.012	0.009	0.017	0.021	0.033
13	0.056	0.044	0.051	0.088	0.194	0.218	0.416	—
14	0.479	0.017	0.017	0.294	0.087	0.138	0.055	0.656
15	−0.040	0.038	0.046	0.067	0.132	0.182	0.292	1.345
16	0.346	0.021	0.021	0.095	0.094	0.182	0.167	—
17	−0.108	0.035	0.048	0.091	0.055	0.213	0.359	1.606
18	0.200	0.006	0.006	0.130	0.024	0.052	0.040	0.133
19	0.037	0.030	0.032	0.085	0.161	0.245	0.248	0.978
20	−0.033	0.044	0.051	0.102	0.119	0.253	0.436	1.662
21	−0.172	0.035	0.049	0.128	0.056	0.095	0.263	0.684
22	−0.099	0.031	0.038	0.160	0.054	0.204	0.113	1.045
23	0.265	0.006	0.006	0.080	0.037	0.055	0.045	0.127
24	0.114	0.022	0.025	0.102	0.049	0.150	0.157	1.116
25	−0.034	0.039	0.047	0.088	0.089	0.183	0.311	0.450
26	−0.289	0.004	0.009	0.027	0.009	0.005	0.038	0.033
27	0.115	0.031	0.032	0.100	0.141	0.279	0.276	1.533
28	0.207	0.024	0.025	0.187	0.087	0.133	0.118	0.525
29	0.098	0.045	0.048	0.123	0.198	0.385	0.396	1.020
30	0.210	0.031	0.036	0.076	0.083	0.137	0.300	0.664
31	0.213	0.027	0.030	0.080	0.084	0.183	0.351	0.485
32	0.235	0.006	0.006	0.068	0.019	0.057	0.031	0.113
33	0.364	0.026	0.026	0.166	0.140	0.146	0.191	0.733
34	−0.079	0.037	0.045	0.061	0.148	0.322	0.484	1.568
35	−0.135	0.029	0.030	0.094	0.130	0.355	0.238	1.439
36	0.519	0.030	0.030	0.388	0.171	0.244	0.134	1.172
37	−0.038	0.004	0.005	0.008	0.016	0.032	0.034	0.139
38	−0.786	0.004	0.020	0.004	0.011	0.038	0.029	0.129
39	−0.168	0.021	0.023	0.042	0.099	0.197	0.157	—
40	−0.192	0.029	0.035	0.062	0.067	0.174	0.261	—

Table 2. 95 percent confidence intervals for ranks: 40 jurisdictions in TSA

(Width of interval represents sampling error. Ranks are based on state mean scores.)

lower limit	observed rank	upper limit	Observed Rank ||
2	1	1	-01
3	2	1	-02-
4	3	2	-03-
6	4	3	--04-
6	5	4	-05-
6	6	3	06---
8	7	7	-07
10	8	8	--08
12	9	7	---09--
13	10	7	---10---
12	11	8	-11---
12	12	9	12---
17	13	12	----13-
16	14	13	--14-
19	15	12	----15---
21	16	13	-----16---
20	17	14	---17---
21	18	15	---18---
23	19	16	----19---
27	20	17	-------20---
26	21	17	-----21----
24	22	20	--22--
27	23	18	----23-----
26	24	22	--24--
28	25	20	---25-----
29	26	22	---26----
30	27	22	---27-----
32	28	25	----28---
33	29	27	----29--
34	30	26	----30---
33	31	29	--31--
33	32	28	-32----
34	33	28	-33-----
35	34	32	-34--
36	35	34	-35-
36	36	35	36-
37	37	37	37
38	38	38	38
39	39	39	39
40	40	40	40

- represents a difference of 1 rank

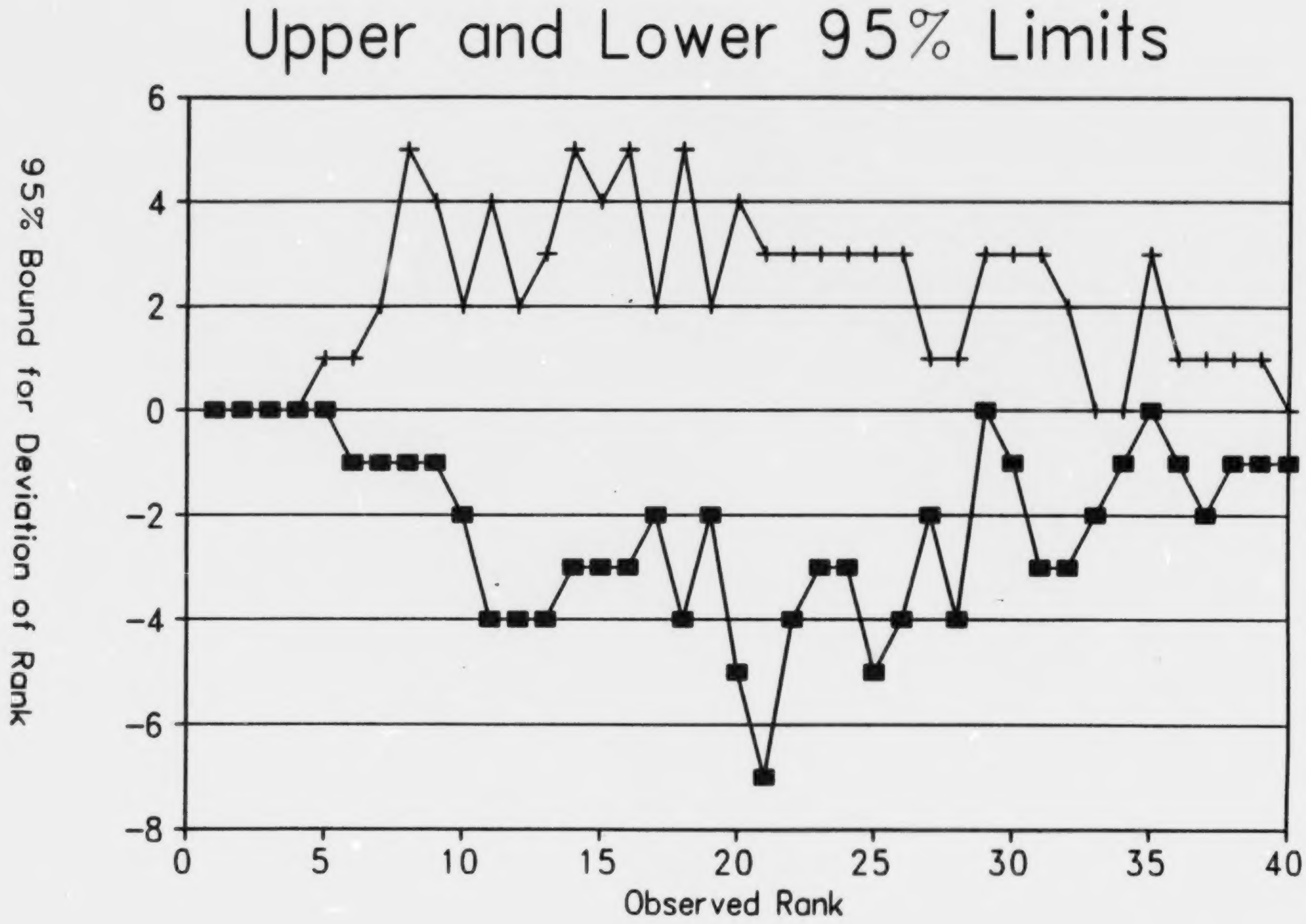

Upper and Lower 95% Limits
95% Bound for Deviation of Rank
6
4
2
0
−2
−4
−6
−8
0
5
10
15
20
25
30
35
40
Observed Rank

Table 3. Coefficients of variation for proportion exceeding basic cutpoint

Standard errors for subgroups were estimated under the assumption that the design effect for the subgroup mean would equal 90 percent of the design effect for the statewide population. The subgroup proportion whose scores exceeded the appropriate proficiency level was assumed to be the same for each subgroup in the state, and therefore equal to the statewide proportion. This assumption was used for the purposes of illustrative calculation of sampling errors. If this assumption overestimated the proportion for a subgroup then, typically, the coefficient of variation for the subgroup would be underestimated; this effect is most pronounced for Tables 4 and 5.

	STATEWIDE		COEFFICIENTS OF VARIATION FOR SUBGROUPS					
	PROP.	CV	WHITE	BLACK	HISPAN	ASIAN	AM.IND.	OTHER
1	0.429	0.029	0.035	0.051	0.128	0.300	0.169	1.474
2	0.495	0.025	0.025	0.134	0.118	0.298	0.159	0.757
3	0.457	0.032	0.038	0.056	0.141	0.293	0.294	0.000
4	0.506	0.007	0.011	0.047	0.010	0.059	0.021	0.113
5	0.749	0.020	0.021	0.069	0.093	0.145	0.186	0.568
6	0.363	0.047	0.061	0.073	0.194	0.457	0.418	2.522
7	0.601	0.027	0.030	0.079	0.113	0.193	0.086	1.334
8	0.073	0.008	0.064	0.009	0.018	0.117	0.110	0.209
9	0.613	0.024	0.026	0.064	0.103	0.169	0.154	0.455
10	0.793	0.013	0.012	0.085	0.057	0.128	0.099	0.365
11	0.241	0.006	0.023	0.050	0.013	0.007	0.061	0.000
12	0.554	0.002	0.002	0.010	0.007	0.014	0.017	0.026
13	0.571	0.026	0.030	0.052	0.115	0.130	0.247	0.000
14	0.840	0.008	0.008	0.147	0.043	0.069	0.028	0.328
15	0.554	0.032	0.039	0.057	0.113	0.155	0.249	1.145
16	0.767	0.011	0.011	0.048	0.047	0.091	0.083	0.000
17	0.530	0.034	0.046	0.088	0.053	0.206	0.347	1.552
18	0.737	0.004	0.004	0.079	0.014	0.032	0.024	0.081
19	0.604	0.021	0.022	0.059	0.112	0.171	0.173	0.682
20	0.582	0.036	0.042	0.084	0.098	0.209	0.360	1.374
21	0.495	0.031	0.044	0.114	0.050	0.085	0.234	0.608
22	0.556	0.026	0.032	0.138	0.047	0.176	0.097	0.901
23	0.738	0.004	0.004	0.053	0.024	0.036	0.030	0.083
24	0.663	0.019	0.021	0.085	0.041	0.125	0.130	0.929
25	0.571	0.032	0.039	0.073	0.074	0.151	0.256	0.371
26	0.438	0.005	0.010	0.030	0.010	0.006	0.042	0.037
27	0.658	0.020	0.021	0.065	0.092	0.183	0.181	1.006
28	0.697	0.018	0.019	0.141	0.066	0.100	0.089	0.396
29	0.639	0.033	0.035	0.090	0.145	0.283	0.291	0.748
30	0.663	0.022	0.025	0.053	0.058	0.096	0.209	0.463
31	0.672	0.019	0.020	0.055	0.057	0.125	0.240	0.331
32	0.738	0.004	0.004	0.046	0.013	0.039	0.021	0.077
33	0.770	0.014	0.014	0.086	0.073	0.076	0.099	0.381
34	0.531	0.028	0.034	0.046	0.111	0.243	0.365	1.182
35	0.509	0.032	0.033	0.102	0.143	0.388	0.261	1.575
36	0.844	0.016	0.016	0.210	0.093	0.132	0.072	0.636
37	0.535	0.003	0.004	0.006	0.013	0.025	0.027	0.110
38	0.194	0.007	0.039	0.008	0.023	0.074	0.057	0.255
39	0.507	0.022	0.025	0.046	0.107	0.214	0.170	0.000
40	0.480	0.029	0.035	0.061	0.066	0.172	0.259	0.000

Table 4. Coefficients of variation for proportion exceeding proficient cutpoint

Standard errors for subgroups were estimated under the assumption that the design effect for the subgroup mean would equal 90 percent of the design effect for the statewide population. The subgroup proportion whose scores exceeded the appropriate proficiency level was assumed to be the same for each subgroup in the state, and therefore equal to the statewide proportion. This assumption was used for the purposes of illustrative calculation of sampling errors. If this assumption overestimated the proportion for a subgroup then, typically, the coefficient of variation for the subgroup would be underestimated; this effect is most pronounced for Tables 4 and 5.

	STATEWIDE		COEFFICIENTS OF VARIATION FOR SUBGROUPS					
	PROP.	CV	WHITE	BLACK	HISPAN	ASIAN	AM.IND.	OTHER
1	0.092	0.078	0.095	0.136	0.347	0.809	0.455	3.978
2	0.096	0.064	0.064	0.344	0.302	0.763	0.408	1.941
3	0.091	0.075	0.088	0.132	0.332	0.690	0.691	0.000
4	0.101	0.011	0.017	0.071	0.016	0.089	0.032	0.172
5	0.246	0.062	0.064	0.210	0.285	0.444	0.569	1.738
6	0.055	0.126	0.161	0.194	0.515	1.210	1.107	6.678
7	0.135	0.076	0.084	0.222	0.318	0.545	0.244	3.760
8	0.006	0.026	0.200	0.028	0.055	0.367	0.346	0.658
9	0.166	0.065	0.070	0.169	0.272	0.448	0.408	1.204
10	0.278	0.045	0.045	0.306	0.207	0.464	0.358	1.320
11	0.039	0.011	0.040	0.089	0.024	0.012	0.109	0.000
12	0.151	0.006	0.006	0.025	0.019	0.035	0.044	0.067
13	0.192	0.082	0.095	0.163	0.360	0.406	0.773	0.000
14	0.300	0.021	0.022	0.379	0.112	0.178	0.072	0.847
15	0.171	0.070	0.086	0.125	0.245	0.337	0.542	2.491
16	0.256	0.037	0.037	0.164	0.162	0.313	0.288	0.000
17	0.125	0.058	0.080	0.152	0.092	0.357	0.600	2.687
18	0.192	0.023	0.023	0.478	0.088	0.193	0.146	0.491
19	0.159	0.062	0.065	0.174	0.330	0.503	0.509	2.011
20	0.149	0.079	0.092	0.183	0.214	0.455	0.783	2.983
21	0.139	0.081	0.116	0.299	0.131	0.223	0.616	1.602
22	0.127	0.053	0.066	0.279	0.094	0.356	0.197	1.822
23	0.215	0.015	0.014	0.192	0.088	0.131	0.107	0.303
24	0.184	0.046	0.052	0.212	0.102	0.311	0.325	2.316
25	0.164	0.060	0.074	0.138	0.140	0.286	0.487	0.706
26	0.123	0.007	0.016	0.046	0.015	0.009	0.066	0.058
27	0.179	0.069	0.071	0.219	0.309	0.612	0.605	3.365
28	0.232	0.047	0.048	0.361	0.168	0.257	0.227	1.013
29	0.190	0.069	0.073	0.188	0.303	0.589	0.606	1.560
30	0.227	0.046	0.054	0.113	0.123	0.205	0.448	0.992
31	0.231	0.040	0.044	0.119	0.124	0.272	0.520	0.718
32	0.204	0.020	0.021	0.229	0.064	0.193	0.105	0.381
33	0.246	0.041	0.041	0.256	0.216	0.226	0.296	1.133
34	0.150	0.079	0.097	0.130	0.315	0.689	1.034	3.350
35	0.104	0.079	0.081	0.256	0.356	0.970	0.651	3.932
36	0.313	0.041	0.041	0.531	0.234	0.334	0.183	1.604
37	0.160	0.007	0.008	0.013	0.027	0.055	0.059	0.239
38	0.026	0.030	0.157	0.031	0.090	0.298	0.229	1.021
39	0.103	0.066	0.074	0.135	0.316	0.632	0.503	0.000
40	0.127	0.069	0.084	0.148	0.159	0.414	0.622	0.000

Table 5. Coefficients of variation for proportion exceeding advanced cutpoint

Standard errors for subgroups were estimated under the assumption that the design effect for the subgroup mean would equal 90 percent of the design effect for the statewide population. The subgroup proportion whose scores exceeded the appropriate proficiency level was assumed to be the same for each subgroup in the state, and therefore equal to the statewide proportion. This assumption was used for the purposes of illustrative calculation of sampling errors. If this assumption overestimated the proportion for a subgroup then, typically, the coefficient of variation for the subgroup would be underestimated; this effect is most pronounced for Tables 4 and 5.

	STATEWIDE		COEFFICIENTS OF VARIATION FOR SUBGROUPS					
	PROP.	CV	WHITE	BLACK	HISPAN	ASIAN	AM.IND.	OTHER
1	0.002	0.414	0.499	0.720	1.829	4.272	2.403	20.995
2	0.001	0.578	0.578	3.111	2.737	6.908	3.691	17.569
3	0.006	0.280	0.331	0.493	1.239	2.577	2.582	0.000
4	0.003	0.021	0.032	0.134	0.030	0.169	0.060	0.327
5	0.011	0.244	0.252	0.825	1.119	1.743	2.234	6.828
6	0.002	0.385	0.494	0.597	1.579	3.712	3.396	20.492
7	0.003	0.383	0.423	1.115	1.595	2.738	1.223	18.882
8	0.001	0.032	0.247	0.035	0.068	0.453	0.427	0.813
9	0.008	0.215	0.232	0.561	0.905	1.488	1.356	4.002
10	0.019	0.152	0.150	1.024	0.695	1.556	1.200	4.426
11	0.004	0.018	0.065	0.145	0.039	0.020	0.176	0.000
12	0.008	0.023	0.024	0.104	0.077	0.145	0.183	0.278
13	0.023	0.276	0.318	0.546	1.206	1.359	2.587	0.000
14	0.016	0.043	0.044	0.767	0.226	0.360	0.145	1.712
15	0.013	0.235	0.289	0.421	0.828	1.135	1.827	8.404
16	0.018	0.104	0.105	0.466	0.459	0.889	0.816	0.000
17	0.006	0.274	0.379	0.715	0.432	1.682	2.831	12.677
18	0.008	0.015	0.015	0.304	0.056	0.122	0.093	0.312
19	0.006	0.253	0.264	0.708	1.341	2.044	2.069	8.168
20	0.009	0.307	0.355	0.711	0.829	1.763	3.035	11.564
21	0.005	0.232	0.330	0.853	0.373	0.634	1.755	4.565
22	0.007	0.181	0.223	0.951	0.320	1.213	0.669	6.200
23	0.012	0.052	0.050	0.681	0.313	0.466	0.381	1.076
24	0.005	0.204	0.227	0.931	0.450	1.369	1.431	10.198
25	0.013	0.130	0.160	0.298	0.302	0.617	1.050	1.521
26	0.009	0.030	0.067	0.196	0.065	0.036	0.278	0.245
27	0.011	0.275	0.284	0.877	1.238	2.450	2.424	13.479
28	0.012	0.169	0.175	1.315	0.610	0.933	0.827	3.685
29	0.009	0.271	0.286	0.741	1.193	2.316	2.382	6.135
30	0.018	0.137	0.160	0.335	0.364	0.605	1.323	2.931
31	0.018	0.112	0.121	0.329	0.344	0.751	1.439	1.987
32	0.007	0.082	0.084	0.920	0.258	0.777	0.421	1.534
33	0.018	0.132	0.132	0.825	0.697	0.728	0.952	3.652
34	0.009	0.244	0.300	0.401	0.977	2.134	3.203	10.381
35	0.002	0.408	0.419	1.321	1.838	5.008	3.360	20.302
36	0.021	0.093	0.093	1.197	0.528	0.754	0.412	3.618
37	0.011	0.013	0.015	0.025	0.052	0.104	0.111	0.454
38	0.004	0.025	0.130	0.026	0.075	0.248	0.191	0.848
39	0.002	0.330	0.369	0.675	1.576	3.151	2.504	0.000
40	0.005	0.271	0.332	0.582	0.628	1.631	2.453	0.000

Summary and Recommendations

The sample design was the standard kind of design used in high quality major national education surveys. It seems to have been executed with care and due attention to details. It appears that the sample might be strenthened in two respects. These are not major points, but they should lead to increased accuracy. First, it is possible that the sample could be made more efficient through improved stratification of schools. *In future assessments, attention should be give to stratifying schools within a state by test scores on statewide exams.* Second, transfer students seem to be markedly underrepresented in the survey. *More attention should be given to capturing transfer students in the assessment.*

Participation rates were reasonable for a survey of this nature. The overall participation rate for schools was about 94 percent, although it was less than 80 percent in two states, Illinois (78 percent) and Oklahoma (76 percent). The state average participation rates of the selected students was between 92 percent and 96 percent, except in two states, where it was in the 80s. *In future assessments, some attention should be given to estimating the amount of systematic error caused by nonparticipation by schools and students within schools. Special attention should also be given to developing additional sources of data to adjust for school and student nonparticipation.*

References

Johnson, E.G., Mislevy, R. J., and Zwick, R. 1990. "Estimation of the Standard Errors of the Adjusted 1986 NAEP Results." *The Effect of Changes in the National Assessment: Disentangling the NAEP 1985-86 Reading Anomaly*. Ed. A. E. Beaton and R. Zwick. Princeton, NJ: Educational Testing Service.

Madow, W. G., Nisselson, H., and Olkin, I. 1983. *Incomplete Data in Sample Surveys, Volume 1, Report and Case Studies*. New York: Academic Press.

Rust, K. April 1991. "The 1990 Trial St[illegible] Assessment: Sampling, Weighting, and Participation." Paper Presented [illegible] 1991 Annual Meeting of American Educational Research Association, Chicago, IL.

Spencer, B. D. 1991. "Eligibility/Exclusion Issues in the 1990 Trial State Assessment." (in this volume.)

Westat, Inc. 1990. *Report on Field Operations and Data Collection Activities, 1990 National and Trial State Assessments* August 20, 1990 Draft, section 3.8.2. Rockville, MD.

Eligibility/Exclusion Issues in the 1990 Trial State Assessment

Bruce D. Spencer
Methodology Research Center, NORC
Northwestern University

November 1991

Acknowledgements

The author is grateful to Jackie Belanger, George Bohrnstedt, Bob Groves, Lyle Jones, Thomas Kerins, Bob Linn, Keith Rust, and Lorrie Shepard for their generous comments and assistance.

Introduction

The 1990 Trial State Assessment (TSA) was restricted to the eighth grade public school population in 37 states, Washington, DC, and 2 U.S. territories. The survey excluded certain large groups of students, some by design (e.g., private school students and students with limited English proficiency or individualized education plans) and some by chance (nonparticipants, movers who "fell through the cracks" of the survey procedure, and absentees). This report discusses issues related to the exclusions by design. Another report (Spencer 1991) discusses the effects of nonparticipation by students who were intended to participate in the TSA, along with other sampling issues.

The study was charged with performing several tasks, including

1. to review rates of exclusions and nonparticipation (by cause) in TSA,
2. to consider possible effects of variable exclusion criteria across or within states,
3. to consider differences between the design for exclusion and the actual field practice for exclusion,
4. to propose ongoing data collection features embedded in state assessments to permit statistical evaluation of the impact of exclusions and nonparticipation,
5. to review past research on impact of exclusions on TSA statistics,
6. to review past research on impact of nonparticipation on TSA statistics,
7. to propose future experimental projects to measure impact of exclusions and nonparticipation, and
8. to evaluate contractor procedures to assess bias due to nonparticipation.

Tasks 1-5 were conducted and are described in this report. The issues of nonparticipation (tasks 6-8) are discussed in the report on sampling issues (Spencer 1991) rather than in this report. The reason for discussing them there rather than here is that nonparticipation represents unintentional exclusion, and the techniques for adjusting for such nonparticipation are part of sampling methodology. This report focusses on deliberate, intentional exclusion and its effects.

Who Was Supposed to Be Assessed?

The TSA was designed to demonstrate the feasibility of conducting a statewide assessment in a consistent and controlled manner across states. The TSA covered such a large number of states and schools within states that the number and diversity of jurisdictions covered is more than adequate for demonstration purposes.[1]

However, the design of the TSA allowed for the exclusions of three groups of students. Students enrolled in private schools were excluded, as were students with limited English proficiency (LEP) or individualized education plans (IEP) who were judged to be incapable of taking the test. (IEP students sometimes are referred to as "special education students.") Nationwide, one in six students belongs to at least one of these three groups. Differences among the states in the proportions of students in these groups could have important effects on state-by-state and state-to-national NAEP comparisons. For example, in national NAEP, which tests both private and public school students, private school students tend to perform better than public school students. Because comparisons of states' performances on NAEP inevitably will be made, the exclusion of private school students in the TSA is cause for concern. The rankings of some states conceivably will change markedly with inclusion of the private school students.

Private School Students

Private school students typically made up the largest excluded group in each state. Nationally, about 12 percent of the eighth grade students are enrolled in private schools, but this percentage varies widely across states. Table 1 shows that in seven states fewer than 5 percent of the eighth grade students are in private schools, and in seven others more than 18 percent are in private schools. Wyoming and Utah each enroll only about 2 percent in private schools, whereas Hawaii and the District of Columbia each enroll about 20 percent.

---

[1] The presumption is that all 50 states would participate in a complete statewide assessment.

Table 1. Percent of eighth grade students in private schools, 1987-88

State or Jurisdiction	Percent Private
Alabama	9.4
Alaska	4.7
Arizona	6.3
Arkansas	3.3
California	12.1
Colorado	5.9
Connecticut	14.5
Delaware	18.4
District of Columbia	21.0
Florida	11.8
Georgia	7.2
Hawaii	19.7
Idaho	3.2
Illinois	16.4
Indiana	10.0
Iowa	9.3
Kansas	7.8
Kentucky	10.0
Louisiana	19.4
Maine	5.0
Maryland	15.5
Massachusetts	13.2
Michigan	11.8
Minnesota	10.3
Mississippi	9.0
Missouri	13.6
Montana	5.5
Nebraska	11.5
Nevada	5.2
New Hampshire	8.3
New Jersey	16.2
New Mexico	7.6
New York	17.8
North Carolina	4.7
North Dakota	6.4
Ohio	12.6
Oklahoma	2.9
Oregon	5.8
Pennsylvania	18.4
Rhode Island	18.4
South Carolina	7.3
South Dakota	5.4
Tennessee	7.1
Texas	7.0
Utah	1.6
Vermont	4.9
Virginia	7.1
Washington	7.9
West Virginia	3.6
Wisconsin	17.9
Wyoming	2.0

Data Source: Sampling Frame for NELS:88; see B. D. Spencer, M. R. Frankel, S. J. Ingels, K. A. Rasinski, and R. Tourangeau. 1990. *National Education Longitudinal Study of 1988 Base Year Sample Design Report.* Data Series DR-NELS:88 - 88 - 1.7 Washington, DC: National Center for Education Statistics.

BEST COPY AVAILABLE

The magnitude of the effect of excluding private school students depends on the number of students enrolled in private schools in a state and on the size of the differences between public and private school students' performances on the NAEP items. Since the private school students' performances were not collected for these particular assessment items, we conducted illustrative sensitivity analyses to suggest how large the effects might be. Two sets of analyses were run: the first set was based on state mean scores and the second, on the proportion of students in each state scoring at or above the three thresholds corresponding to basic, proficient, or advanced performance. The proportion of eighth grade students in each state who were enrolled in private schools was estimated from data compiled for the development of the NELS:88 sample by National Opinion Research Center. (See Spencer et al. 1990.)

The analyses based on mean scores will be discussed first.
One sensitivity analysis was based on the simple assumption that,

(Model I) had the private school students been tested, on average their scores would have been 0.6 standard units[2] higher than the scores of their public school counterparts in the state.

This analysis suggests that the rankings for several of the 38 "states" (including the District of Columbia) could change up or down by three or more (Tables 2-3). As Table 3 shows, in the sensitivity analysis, 1 state's rank changed by five, 2 changed by three, 7 changed by two, 13 changed by one, and 15 remained unchanged. The average (across the 38 states) change in rank was 1.0; the average change in mean standard score was 0.06.

The assumption that the difference between the mean scores of public and private students is constant (.6) in each state likely understates the real variation in overall state mean scores. Therefore a second sensitivity analysis was run in which

---

[2] Standard units refer to the score distribution for public school students assessed in the TSA. The difference of 0.6 standard units is consistent with NELS:88 data on eighth grade students.

Table 2. **Mean scores of states under alternative assumptions about private school students and IEP/LEP students**

		Model					
St.	observed	I	II	III	VII	VIII	IX
1	−0.32	−0.29	−0.30	−0.28	−0.35	−0.44	−0.42
2	−0.18	−0.16	−0.17	−0.15	−0.23	−0.39	−0.38
3	−0.24	−0.19	−0.15	−0.10	−0.29	−0.44	−0.35
4	−0.17	−0.13	−0.12	−0.09	−0.21	−0.42	−0.37
5	0.34	0.45	0.51	0.60	0.29	0.16	0.33
6	−0.42	−0.30	−0.21	−0.12	−0.45	−0.56	−0.35
7	0.01	0.03	0.04	0.05	−0.04	−0.21	−0.18
9	0.05	0.13	0.08	0.14	0.02	−0.11	−0.09
10	0.42	0.48	0.52	0.56	0.38	0.26	0.36
12	−0.06	0.05	0.06	0.15	−0.13	−0.32	−0.20
13	0.06	0.10	0.06	0.10	0.01	−0.17	−0.16
14	0.48	0.51	0.49	0.51	0.45	0.37	0.38
15	−0.04	0.05	0.04	0.11	−0.09	−0.23	−0.15
16	0.35	0.41	0.41	0.47	0.31	0.22	0.29
17	−0.11	−0.07	−0.03	0.00	−0.17	−0.36	−0.28
18	0.20	0.22	0.20	0.22	0.17	0.10	0.11
19	0.04	0.11	0.16	0.22	−0.02	−0.18	−0.06
20	−0.03	0.06	0.04	0.13	−0.08	−0.24	−0.16
21	−0.17	−0.10	−0.07	−0.01	−0.25	−0.46	−0.36
22	−0.10	−0.06	−0.09	−0.06	−0.16	−0.30	−0.29
23	0.26	0.31	0.31	0.35	0.22	0.08	0.12
24	0.11	0.15	0.13	0.16	0.07	−0.07	−0.06
25	−0.03	0.07	0.16	0.25	−0.12	−0.29	−0.10
26	−0.29	−0.17	−0.19	−0.10	−0.35	−0.46	−0.37
27	0.11	0.17	0.20	0.25	0.07	−0.08	0.00
28	0.21	0.24	0.25	0.28	0.17	0.09	0.13
29	0.10	0.21	0.20	0.30	0.05	−0.12	−0.01
30	0.21	0.31	0.35	0.43	0.14	−0.09	0.05
31	0.21	0.30	0.35	0.42	0.16	−0.06	0.07
32	0.23	0.25	0.24	0.25	0.19	0.08	0.09
33	0.36	0.43	0.41	0.46	0.33	0.24	0.28
34	−0.08	−0.04	−0.05	−0.02	−0.11	−0.21	−0.18
35	−0.13	−0.07	−0.12	−0.07	−0.18	−0.32	−0.31
36	0.52	0.56	0.57	0.60	0.49	0.40	0.45
37	−0.04	0.07	−0.00	0.09	−0.09	−0.20	−0.16
38	−0.79	−0.66	−0.56	−0.46	−0.81	−0.96	−0.74
39	−0.17	−0.15	−0.15	−0.14	−0.26	−0.45	−0.44
40	−0.19	−0.12	−0.13	−0.07	−0.25	−0.44	−0.37
		AVERAGE DIFFERENCE FROM OBSERVED					
		0.06	0.07	0.13	0.04	0.18	0.12

Table 3. **Mean-score based rankings of states under alternative assumptions about private school students and IEP/LEP students**

		Model					
St.	observed	I	II	III	VII	VIII	IX
1	36	0	1	1	0	−5	0
2	32	1	2	4	−2	−3	3
3	34	1	−2	−1	0	−1	−5
4	31	0	−1	0	−2	−1	2
5	6	−2	−3	−4	0	0	−2
6	37	0	−1	−3	0	0	−7
7	19	5	3	5	0	1	4
9	17	−1	1	2	−1	−2	0
10	3	0	−1	0	0	0	0
12	24	−1	−4	−6	1	3	1
13	16	2	3	6	1	1	4
14	2	0	2	2	0	0	0
15	23	−1	0	−2	−1	−1	−4
16	5	1	0	0	0	0	0
17	27	0	−2	−2	0	1	−1
18	12	0	0	4	−3	−5	−3
19	18	−1	−3	−3	0	0	−3
20	20	1	1	0	0	3	2
21	30	−1	−3	−4	2	6	1
22	26	0	2	2	0	−1	1
23	7	0	2	2	0	3	1
24	14	1	3	3	0	−2	2
25	21	−2	−5	−7	3	3	−3
26	35	−1	0	−3	0	0	−3
27	13	1	1	0	0	0	0
28	11	0	−1	0	−1	−3	−4
29	15	−2	−2	−5	0	1	−1
30	10	−2	−3	−3	2	4	2
31	9	0	−1	−1	2	2	2
32	8	2	3	4	0	1	2
33	4	1	2	2	0	0	2
34	25	0	1	2	−2	−4	−1
35	28	0	1	2	0	−2	0
36	1	0	0	0	0	0	0
37	22	−2	2	1	−1	−3	−1
38	38	0	0	0	0	0	0
39	29	3	4	6	4	5	8
40	33	−3	−2	−4	−2	−1	1
		AVERAGE DIFFERENCE FROM OBSERVED					
		1.00	1.79	2.53	0.75	1.70	1.90

(Model II) the difference between the public and private mean score in each state was assumed to be drawn at random uniformly between 0.1 and 1.1 (with a mean of .6).

In the sensitivity analysis, 1 state's rank changed by seven ranks, 3 changed by six, 2 changed by five, 6 changed by four, 5 changed by three, 9 changed by two, 4 changed by one, and 8 remained unchanged (Table 3). The average change in rank increased to 1.8. The average change in mean standard score was 0.07 (Table 2). (These results would change if the simulation were repeated, since the scores were randomly generated.)

A third sensitivity analysis assumed

(Model III) the mean difference was 1.1 rather than .6, and the difference between the public and private mean scores in each state was assumed to be drawn at random uniformly between 0.6 and 1.6. (with a mean of 1.1).

In this simulation, 1 state's rank changed by five ranks, 2 changed by four, 8 changed by three, 10 changed by two, 11 changed by one, and 6 remained unchanged (Table 3). The average change in rank increased to 2.5, and the average change in mean standard score was 0.13 (Table 2).

In each simulation, the effect of excluding the private school students is at least twice as large as the effect of sampling error in the TSA (see Spencer 1991).

The analyses based on proportions of students meeting basic, proficient, or advanced performance standards will now be discussed. The analyses were based on the logits of the proportions, with logit(p) defined as the logarithm of p/(1-p) for a proportion p. National estimates of the difference between the logits of the proportions of public school students and private school students meeting the various performance standards were developed, and these differences were assumed to describe state-level differences, more or less (Table 4).

Table 4. Proportions of public and private students assumed to meet performance standards

	Performance Level		
Students	Basic	Proficient	Advanced
Public	60.1	17.1	1.0
Private	78.1	26.7	1.1

The meaning of "more or less" is made precise in Models IV - VI below. The assumption is only a crude approximation, but it does give insights into the sensitivity of the state proportions to exclusion of private school students.

The first sensitivity analyses assumed that

(Model IV) the same national public-private logit differences were present in each state as well.

This analysis suggested that the percentages meeting the basic and proficient standards would have increased by about 2 percent and 1 percent respectively if private students had been tested (Table 5). The average change in rank would have been about one (for each standard), with many states' ranks not changing but one state changing in rank by as much as four (Table 6). The changes in proportions meeting the advanced standards were minor; only two states changed rank at all and only by one.

The second set of sensitivity analyses assumed that

(Model V) the difference between the logits for public and private students for a state was chosen at random from a uniform distribution ranging up or down 0.5 from the national difference in logits.

(This assumption parallels Model II for state mean scores, as described previously.) This analysis too suggested that the percentages meeting the basic and proficient standards would have increased by about 2 percent and 1 percent respectively if

Table 5. Percentage of students scoring at or above basic, proficient, or advanced thresholds under alternative assumptions about private school students

	Basic				Proficient				Advanced			
St.	obsvd	IV	V	VI	obsvd	IV	V	VI	obsvd	IV	V	VI
1	42.9	43.9	44.0	43.7	9.2	9.5	9.6	9.5	.2	.2	.2	.2
2	49.5	50.2	50.5	49.5	9.6	9.9	10.1	9.6	.1	.1	.1	.1
3	45.7	47.6	46.7	46.5	9.1	9.7	9.2	9.2	.6	.6	.6	.6
4	50.6	52.1	52.0	52.3	10.1	10.6	10.6	10.8	.3	.3	.2	.3
5	74.9	77.2	76.4	76.2	24.6	26.7	25.3	25.0	1.1	1.2	1.2	1.1
6	36.3	40.4	38.2	39.9	5.5	6.3	5.6	6.1	.2	.2	.3	.2
7	60.1	60.6	60.4	60.9	13.5	13.7	13.5	14.0	.3	.3	.3	.3
9	61.3	63.4	64.0	64.6	16.6	17.7	18.6	19.9	.8	.8	.8	.8
10	79.3	80.3	79.9	80.7	27.8	28.9	28.1	30.5	1.9	1.9	1.9	2.2
12	55.4	58.9	58.8	59.0	15.1	16.7	16.6	16.8	.8	.8	.8	.9
13	57.1	58.4	59.0	58.6	19.2	19.9	20.7	20.1	2.3	2.3	2.2	2.2
14	84.0	84.5	84.6	84.1	30.0	30.7	31.4	29.8	1.6	1.6	1.6	1.6
15	55.4	58.3	58.7	57.5	17.1	18.6	18.9	17.8	1.3	1.4	1.3	1.2
16	76.7	78.0	78.1	76.8	25.6	27.0	27.1	25.2	1.8	1.9	2.0	1.7
17	53.0	54.4	53.7	55.2	12.5	13.1	12.6	14.0	.6	.6	.6	.6
18	73.7	74.2	74.3	74.3	19.2	19.6	19.9	19.8	.8	.8	.8	.8
19	60.4	62.6	61.8	63.9	15.9	17.0	16.2	19.0	.6	.6	.6	.5
20	58.2	61.3	61.6	59.8	14.9	16.4	16.7	15.2	.9	.9	.9	.8
21	49.5	52.0	51.4	53.5	13.9	14.9	14.5	16.6	.5	.5	.5	.5
22	55.6	56.8	57.3	56.7	12.7	13.2	13.7	13.1	.7	.7	.7	.7
23	73.8	74.9	75.0	75.4	21.5	22.5	22.5	23.7	1.2	1.2	1.3	1.3
24	66.3	67.3	67.6	66.4	18.4	19.0	19.5	18.2	.5	.5	.5	.5
25	57.1	60.5	58.8	57.1	16.4	18.1	16.7	15.8	1.3	1.4	1.4	1.3
26	43.8	47.9	48.5	43.3	12.3	13.8	14.2	11.6	.9	.9	1.0	.8
27	65.8	67.4	67.0	68.0	17.9	18.9	18.4	19.9	1.1	1.1	1.2	1.1
28	69.7	70.6	70.4	71.1	23.2	23.9	23.6	25.2	1.2	1.2	1.2	1.3
29	63.9	67.0	67.1	66.7	19.1	21.0	21.1	20.5	.9	1.0	1.0	.8
30	66.3	68.9	68.2	68.8	22.7	24.6	23.7	24.5	1.8	1.8	1.8	2.0
31	67.2	69.5	68.7	70.4	23.1	24.7	23.7	26.6	1.8	1.8	1.8	2.0
32	73.8	74.1	74.2	74.1	20.4	20.6	20.7	20.6	.7	.7	.7	.7
33	77.0	78.2	78.4	78.8	24.6	25.8	26.2	27.7	1.8	1.8	1.8	1.9
34	53.1	54.5	54.8	53.4	15.0	15.6	15.9	14.9	.9	.9	.9	1.0
35	50.9	52.9	53.8	52.5	10.4	11.0	11.8	10.7	.2	.2	.2	.2
36	84.4	84.9	84.8	84.8	31.3	32.1	31.8	31.6	2.1	2.1	2.1	2.1
37	53.5	57.1	58.4	54.4	16.0	17.7	19.2	15.8	1.1	1.1	1.0	1.1
38	19.4	22.9	20.8	21.2	2.6	3.0	2.7	2.7	.4	.4	.4	.4
39	50.7	51.4	51.5	51.5	10.3	10.5	10.5	10.6	.2	.2	.2	.2
40	48.0	50.4	50.5	49.2	12.7	13.6	13.6	12.8	.5	.6	.5	.6
AVERAGE DIFFERENCE FROM OBSERVED												
		2.0	1.7	1.5		1.0	.9	1.1		.0	.0	.1

Table 6. **Rankings of states based on percentage of students scoring at or above basic, proficient, or advanced thresholds under alternative assumptions about private school students**

	Basic				Proficient				Advanced			
	Rank	Difference in Rank			Rank	Difference in Rank			Rank	Difference in Rank		
St.	obsvd	IV	V	VI	obsvd	IV	V	VI	obsvd	IV	V	VI
1	36	0	0	-1	35	1	0	0	37	0	0	0
2	32	1	1	0	34	0	0	0	38	0	0	0
3	34	1	1	0	36	-1	0	0	25	0	0	0
4	30	-1	-1	0	33	-1	-1	-2	33	0	1	0
5	6	0	0	0	6	-1	0	2	13	0	1	0
6	37	0	0	0	37	0	0	0	34	0	-1	2
7	18	1	1	0	26	1	3	0	32	0	0	0
9	16	0	0	0	18	2	0	-3	20	0	0	2
10	3	0	0	0	3	0	0	-1	3	0	1	-1
12	24	-3	-3	-4	22	0	0	-2	21	0	0	-4
13	21	1	-1	0	13	0	0	0	1	0	0	0
14	2	0	0	0	2	0	0	1	8	0	0	0
15	23	0	0	-1	17	0	0	2	9	0	1	3
16	5	0	0	0	4	0	0	2	4	0	-1	3
17	27	0	1	-2	29	1	1	-2	26	0	0	0
18	9	-1	-1	-1	12	2	2	4	22	0	0	-1
19	17	0	0	0	21	0	2	-4	27	0	0	1
20	19	-1	-1	0	24	-1	-4	0	17	0	2	1
21	31	-1	0	-4	25	0	0	-4	29	0	0	1
22	22	3	3	2	27	2	0	1	24	0	0	-1
23	7	0	0	0	10	0	0	0	12	0	-1	-3
24	12	2	1	3	15	0	0	3	30	0	0	-1
25	20	0	2	3	19	-1	2	4	10	0	-1	1
26	35	-1	-1	1	30	-4	-4	0	19	-1	-2	1
27	14	-1	1	-1	16	0	3	-2	14	0	-2	1
28	10	0	0	0	7	2	2	0	11	0	2	-1
29	15	0	-1	-1	14	-3	-3	-2	16	0	0	3
30	13	-1	-1	-1	9	-1	-1	0	7	0	0	-2
31	11	0	0	0	8	-1	-1	-3	6	0	0	-2
32	8	1	1	1	11	1	1	0	23	0	0	1
33	4	0	0	0	5	1	0	-1	5	0	0	1
34	26	0	0	2	23	1	1	2	18	1	0	-2
35	28	0	-1	1	31	0	0	1	36	0	-1	-2
36	1	0	0	0	1	0	0	0	2	0	0	1
37	25	-1	-1	1	20	-1	-4	2	15	0	0	-1
38	38	0	0	0	38	0	0	0	31	0	0	0
39	29	2	1	2	32	1	1	1	35	0	1	0
40	33	-1	-1	0	28	0	0	1	28	0	0	-1
AVERAGE DIFFERENCE FROM OBSERVED												
		.63	.68	.84		.79	.95	1.37		.05	.47	1.16

private students had been tested (Table 5). The average change in rank would have been about one (for each standard), with many states' ranks not changing but several states changing in rank by as much as four (Table 6). The changes in proportions meeting the advanced standards were a little larger than the preceding analysis, in that several states changed rank by three or four. The average change in rank was about one.

The third set of sensitivity analyses was similar to the second except that the range of the uniform distribution was set to be twice as large, i.e., plus or minus one. Specifically,

(Model VI) the difference between the logits for public and private students for a state was chosen at random from a uniform distribution ranging up or down 1.0 from the national difference in logits.

The conclusions are similar to those based on Model V (Tables 5 and 6).

In all, the magnitude of the effect of exclusion of private school students ranges from about the same (proportions meeting performance standards) to twice as large (for state mean scores) as sampling error. *Therefore, it is recommended that state NAEP include private as well as public school students.*

Limited English Proficiency (LEP) and Individualized Education Plan (IEP) Exclusions

The exclusion criteria for LEP and IEP students were part of the sampling design and were implemented locally. The local administrators were told to exclude students who could not be expected to be able to take the test, but to include doubtful cases in the assessment.

The criteria of excluding students from the assessment were as follows (Westat 1990):

1. **The intent is to assess all selected students. Therefore, all selected students who are capable of participating in the assessment should be assessed.**

2. Some of the students identified on the Administration Schedule as LEP or having an IEP may be incapable of participating meaningfully in the assessment. The Local Administrator, with the advice of staff members knowledgeable about the IEP/LEP students, may exclude such students, as described below.

3. A student identified on the Administration Schedule as LEP may be excluded from the assessment if he/she

   - Is a native speaker of a language other than English; AND
   - Has been enrolled in an English-speaking school (not including a bilingual education program) for less than two years; AND
   - Is judged to be incapable of taking part in the assessment.

4. A student identified on the Administration Schedule as having an IEP or equivalent classification may be excluded form the assessment if:

   - The student is mainstreamed less than 50 percent of the time in academic subjects and is judged incapable of participating meaningfully in the assessment,

   OR

   - The IEP team or equivalent group has determined that the student is incapable of participating meaningfully in the assessment.

5. IEP/LEP students meeting the above criteria should be assessed if, in the judgment of school staff, they are capable of taking the assessment. **WHEN THERE IS DOUBT, INCLUDE THE STUDENT.**

Overall, about 1 percent of the students were excluded for reasons of limited English proficiency and about 4 percent because they had individualized education plans. But the percentages across states varied, with LEP exclusions ranging from near 0 percent in West Virginia and about 2 percent in New Jersey, New York, Rhode Island, and Texas, to 5 percent in California. IEP exclusions ranged from about 2 percent in Montana to 8 percent in Arkansas.

Attempts to compare the proportions of students classified as IEP and LEP in the TSA with other available data are hindered by the weaknesses of the available comparison data. First, consider national statistics on the number of IEP and LEP students, which were available from NCES but only for the aggregate of ages 5-17 (LEP) or 6-17 (IEP). The differences in the percentages of LEP students tended to be about 1 percent lower for the TSA than for the comparison statistics, except that the comparison data put the percentage of LEP students in New Mexico at 16.4, whereas the TSA put it somewhat under 2 percent.[3] It is not clear whether these discrepancies in proportions of LEP students are due to imperfections in the comparison data, imperfections in the TSA data, or differences in the age groups. It is quite plausible, however, that the definition of LEP was applied inconsistently within states as well as across states. This is not necessarily a flaw in the implementation of NAEP, since local administrators were not instructed to apply objective criteria to decide who should be excluded as LEP. The problem is that the operationalization of the concept of "LEP" was not fully specified in the design.

The discrepancies for IEP for the 37 states and the District of Columbia ranged from about 3.5 percent lower to 3.5 percent higher in the TSA, averaging about 1 percent lower in the TSA; see Table 7.[4] The correlation between the two sets of rates was about the same as between the two sets of LEP rates, 0.63. It is not clear whether these discrepancies for IEP are due to imperfections in the comparison data, imperfections in the TSA data, or differences in the age groups. The most common handicaps leading to IEP status are, in decreasing prevalence (national percentages in parentheses), learning disability (4.7), speech impairment (2.4), mental

---

[3] The comparison data are for ages 5-17 and were taken from Office of Special Education Programs (undated). The comparison data were available for only 31 of the 40 areas. Except for New Mexico, the discrepancies ranged from 4 percent lower in the TSA to 1 percent higher, with a standard deviation of about 1 percent. The correlation between the two sets of rates was 0.68.

[4] Data are for 1988-89 and were taken from Table AA25 in Office of Special Education Programs (1990). (Proportions include children aged 6-17 served under Chapter 1 of ESEA (SOP) and EMA-O.)

Table 7. Percentages of excluded (IEP and LEP) students in the TSA[1]-- Preliminary data

	Total Percent Excluded Students	Total Percent IEP	Alternative[2] Total Percent IEP	Percent IEP Excluded	Total Percent LEP	Alternative[3] Total Percent LEP	Percent LEP Excluded
Alabama	5.7	9.7	11.56	5.7	0.2	N/A	0.1
Arizona	5.4	7.0	8.40	3.7	5.6	7.6	1.5
Arkansas	8.0	11.1	8.98	7.9	0.4	N/A	0.1
California	8.2	7.0	8.20	3.5	8.9	12.6	4.7
Colorado	4.9	8.9	8.26	3.9	1.3	1.4	0.9
Connecticut	7.1	9.6	11.75	5.9	2.1	2.6	1.2
Delaware	4.5	8.8	11.88	4.0	0.9	N/A	0.5
Florida	7.0	9.0	10.75	5.1	2.8	1.9	1.9
Georgia	3.8	6.6	7.44	3.5	0.3	0.3	0.1
Hawaii	5.0	7.4	6.64	3.7	2.9	4.7	1.3
Idaho	2.5	6.2	7.88	2.3	0.6	1.1	0.2
Illinois	5.6	8.7	11.81	4.9	1.0	2.7	0.8
Indiana	5.3	7.4	10.01	5.1	0.1	0.3	0.1
Iowa	3.9	9.6	10.17	3.9	0.4	0.6	0.0
Kentucky	5.2	7.6	10.03	5.0	0.2	0.2	0.2
Louisiana	4.5	6.3	7.34	4.2	0.3	0.9	0.1
Maryland	5.0	10.0	11.29	4.4	1.3	0.9	0.6
Michigan	4.6	8.0	8.67	4.3	0.5	1.7	0.2
Minnesota	3.2	8.3	9.65	2.8	0.9	1.2	0.3
Montana	2.4	6.5	8.72	2.4	0.2	1.9	0.1
Nebraska	3.1	8.3	10.15	3.0	0.4	0.4	0.1
New Hampshire	4.7	11.9	8.42	4.6	0.3	0.4	0.1
New Jersey	7.7	10.3	13.93	5.7	2.4	2.7	1.9
New Mexico	7.3	8.6	10.21	6.0	1.5	16.4	0.9
New York	7.0	8.7	9.67	4.9	3.7	3.9	2.2
North Carolina	3.4	9.1	9.30	3.3	0.3	0.3	0.1
North Dakota	2.9	7.6	8.94	2.8	0.8	4.9	0.1
Ohio	5.8	8.2	10.04	5.7	0.1	0.5	0.1
Oklahoma	5.7	8.0	9.65	5.4	0.6	1.5	0.4
Oregon	3.0	7.5	9.40	2.6	1.1	1.0	0.4
Pennsylvania	5.5	10.0	10.86	5.3	0.5	N/A	0.4
Rhode Island	7.1	11.6	12.92	5.2	3.7	3.1	2.0
Texas	6.6	8.2	8.53	4.9	5.4	8.7	2.0
Virginia	5.7	8.3	9.22	4.6	1.6	N/A	1.3
Washington DC	6.0	5.4	6.68	4.7	1.3	N/A	1.3
West Virginia	5.9	9.7	11.55	5.9	0.1	N/A	0.0
Wisconsin	4.7	7.7	8.32	4.3	0.7	1.0	0.2
Wyoming	3.8	8.3	9.07	3.7	0.5	1.5	0.2
Guam	4.1	5.4	N/A	3.6	1.5	N/A	0.6
Virgin Islands	3.3	3.8	N/A	3.1	0.3	N/A	0.2

[1] To be excluded, a student was supposed to be IEP or LEP and judged incapable of participating in the assessment. A student reported as both IEP and LEP is counted once in the overall exclusion rate (first column) and separately in the remaining columns.

[2] 1988-89 data; see Table C below.

[3] The comparison data are for ages 5-17 and were taken from U.S. Department of Education, Office of Bilingual and Minority Affairs, *A Summary of State Reports on the Number and Condition of Limited English Proficient Students, 1986-87*.

retardation (1.2), emotional disturbance (0.9), multiple handicaps (0.2), hearing disability (0.1), "other" health impairments (0.1), orthopedic impairments (0.1), and visual handicaps (0.05); see Table 8. Since the major disabilities are those with the "softest" (most unreliable) diagnoses, it is reasonable to assume that the classification criteria varied across states. The discrepancies of the magnitude observed could reflect differences in the ages being compared or they could reflect variable criteria *within states*. Thus, local policies for excluding IEP students varied from district to district. Again, the problem is that the operationalization of the concept of "IEP" was not fully specified in the design (as also was the case for "LEP").

Attempts were made to compare the TSA and national NAEP regarding proportions of IEP and LEP students. Data on proportions of students who were excluded for reasons of LEP or IEP for national NAEP[5] were not available from Educational Testing Service (ETS). However, it is possible to compare the TSA and national NAEP on the proportion of test-takers who were classified as IEP or LEP. The TSA included roughly twice as many test-takers classified as IEP or LEP as national NAEP did: 4.0 percent IEP and 1.4 percent LEP in the TSA compared to 1.7 percent IEP and 0.8 percent LEP in national NAEP[6]. This discrepancy implies that the TSA and national NAEP either (a) employed different criteria for classifying the students as IEP/LEP, or (b) excluded students who were "really" IEP/LEP at rather different rates. Explanation (a) is relatively innocuous in its effects on the interpretation of assessment results overall, since it could be that the same students would be tested in the TSA as in national NAEP, although they would be more likely to be classified as IEP/LEP in the TSA. On the other hand, explanation (b), if valid, would be cause for concern since it implies that the TSA was more stringent in testing IEP/LEP students than national NAEP has been--either the TSA is too stringent or national NAEP is too lax. It may well be that the instructions to local administrators to give the assessment to "doubtful" cases explains the entire discrepancy. For example, a school may have designated students as IEP/LEP for funding purposes and, of these, the school may elect to test two thirds for the TSA and only one third for national NAEP. The proportions of IEP/LEP students who

---

[5] "National NAEP" refers to the subset of the national NAEP sample that is comparable to the TSA sample.

[6] Approximate p-values for the differences between TSA and national NAEP are < 0.001 for IEP and 0.08 for LEP.

Table 8. Percentage of children 6-17 served under chapter 1 of ESEA (SOP) and EHA-B by handicapping condition

Data are for 1988-89 and were taken from Table AA25 in *Twelfth Annual Report to Congress, 1990*, Washington, DC. Office of Special Education Programs. (Proportions include children aged 6-17 served under Chapter 1 of ESEA (SOP) and EMA-O.) (Because of rounding, the sum of the percents of individual handicapping conditions may not equal the total percent of all conditions.)

ALL-All Conditions; LD-Learning Disabled; SI-Speech Impaired; MR-Mentally Retarded; EMD-Emotionally Disturbed; HH-Hard of Hearing and Deaf; MH-Multi-Handicapped; OPI-Orthopedically Impaired; OHI-Other Health Impaired; VH-Visually Handicapped.

STATE	ALL	LD	SI	MR	EMD	HH	MH	OPI	OHI	VH
Alabama	11.56	3.90	3.06	3.39	0.77	0.12	0.12	0.06	0.09	0.05
Arizona	8.40	4.78	1.94	0.64	0.55	0.15	0.18	0.07	0.06	0.04
Arkansas	8.98	4.82	1.49	2.24	0.07	0.12	0.12	0.03	0.05	0.04
California	8.20	4.96	1.96	0.41	0.23	0.13	0.10	0.13	0.24	0.05
Colorado	8.26	4.06	1.39	0.48	1.51	0.13	0.51	0.13	0.00	0.04
Connecticut	11.75	6.35	1.94	0.68	2.30	0.12	0.17	0.05	0.07	0.08
Delaware	11.88	6.82	1.64	1.10	1.64	0.16	0.13	0.21	0.11	0.06
District of Columbia	6.68	3.30	1.15	0.92	0.93	0.04	0.15	0.08	0.05	0.05
Florida	10.75	4.56	3.34	1.26	1.23	0.08	0.00	0.11	0.13	0.04
Georgia	7.44	2.19	1.66	1.84	1.53	0.10	0.00	0.05	0.03	0.04
Hawaii	6.64	3.81	1.24	0.62	0.45	0.13	0.11	0.15	0.08	0.04
Idaho	7.88	4.71	1.46	1.13	0.21	0.12	0.02	0.09	0.11	0.03
Illinois	11.81	5.52	3.11	1.27	1.44	0.16	0.01	0.15	0.09	0.06
Indiana	10.01	3.80	3.65	1.82	0.44	0.11	0.07	0.05	0.01	0.05
Iowa	10.17	4.53	1.94	1.93	1.31	0.14	0.10	0.18	0.00	0.03
Kentucky	10.03	3.21	3.34	2.60	0.43	0.12	0.15	0.06	0.04	0.07
Louisiana	7.34	2.96	2.28	1.10	0.45	0.14	0.09	0.11	0.16	0.05
Maryland	11.29	5.79	3.49	0.65	0.54	0.17	0.35	0.09	0.13	0.09
Michigan	8.67	3.90	2.06	1.01	1.18	0.14	0.09	0.21	0.03	0.04
Minnesota	9.65	4.64	1.91	1.25	1.41	0.18	0.02	0.15	0.05	0.05
Montana	8.72	4.83	2.27	0.64	0.39	0.13	0.18	0.06	0.11	0.10
Nebraska	10.15	4.40	2.78	1.39	0.87	0.16	0.13	0.22	0.13	0.06
New Hampshire	9.42	5.78	1.62	0.50	0.93	0.12	0.13	0.08	0.18	0.06
New Jersey	13.93	7.03	4.55	0.42	1.18	0.11	0.51	0.05	0.04	0.04
New Mexico	10.21	4.91	3.03	0.62	1.08	0.13	0.19	0.17	0.03	0.04
New York	9.67	5.81	0.92	0.70	1.58	0.12	0.30	0.07	0.11	0.05
North Carolina	9.30	4.05	2.17	1.71	0.81	0.16	0.10	0.08	0.18	0.05
North Dakota	8.94	4.26	2.94	1.07	0.34	0.13	0.00	0.08	0.05	0.05
Ohio	10.04	3.94	2.77	2.19	0.40	0.11	0.40	0.18	0.00	0.05
Oklahoma	9.65	4.58	2.64	1.79	0.24	0.10	0.19	0.05	0.02	0.04
Oregon	9.40	5.11	2.47	0.62	0.57	0.21	0.00	0.15	0.19	0.07
Pennsylvania	10.86	4.61	3.09	1.84	1.02	0.16	0.00	0.07	0.00	0.07
Rhode Island	12.92	8.66	2.17	0.61	1.00	0.11	0.05	0.11	0.17	0.05
Texas	8.53	4.82	1.78	0.62	0.69	0.11	0.10	0.10	0.25	0.05
Virginia	9.22	4.73	2.28	1.12	0.74	0.11	0.08	0.06	0.05	0.05
West Virginia	11.55	5.27	3.16	2.22	0.64	0.10	0.00	0.08	0.02	0.06
Wisconsin	8.32	2.79	1.65	0.52	1.23	0.03	2.00	0.05	0.03	0.03
Wyoming	9.07	4.91	2.51	0.57	0.53	0.14	0.00	0.14	0.21	0.05
50 States D.C. & P.R.	9.83	4.71	2.39	1.25	0.89	0.13	0.18	0.11	0.12	0.05

were tested could vary from school to school (and state to state) even if consistent definitions of IEP and LEP were used. It is worth noting that, of the students classified as LEP, the proportions excluded ranged from 0 percent (Iowa) to 100 percent (District of Columbia). The correlation between the proportion of students who were LEP (or IEP) and the proportion of LEP (or IEP) students who were excluded was very low, at about -0.1 (+0.1).

Comparisons of the monitored and unmonitored sessions regarding the proportions of tested students who were classified as IEP or LEP were also made. The differences generally were minor, although in some states the IEP differences reached 2 percent.[7] LEP differences were smaller, with only 1 of the 40 as large as 0.005. (There was no consistent direction to the differences.)

The average score for IEP/LEP students who took the test was lower than the average for the other tested students. If all of the IEP/LEP students had been tested, the mean scores for the 40 jurisdictions would have been lower. Several sensitivity analyses were performed in order to provide order of magnitude estimates for how much lower they might have been. One analysis assumed that

(Model VII) the excluded IEP/LEP students in a state would have scored about the same as the tested IEP/LEP students in the state.

This is an optimistic assumption and implies that the average difference in mean scores for the jurisdictions would have changed by about 0.04 standard units (Table 2). The rankings for three states (based on mean scores) would have changed by as much as three or four, but the average change in rank would have been about 0.75 (Table 3).[8] The percentages of students meeting the three performance standards would have changed by 3 percent (basic), 1 percent (proficient), and 0.0 percent (advanced). The rankings would have changed, on average, by about 1.75, 1.65, and

---

[7] Student's t statistics were computed for differences in proportion of tested students who were classified as IEP for each of the 40 jurisdictions. The absolute value of the t statistic exceeds 2 for seven jurisdictions and exceeds 3.36 for two of them. A Bonferroni test at an $\alpha = 0.051$ significance level finds the latter two significant.

[8] The changes in rankings were as follows: 22 did not change, 6 changed one rank, 7 changed two ranks, 2 changed three ranks, and 1 changed four ranks.

1.05 respectively. These results may be regarded as a lower bound for the differences that actually would have occurred if all of the IEP/LEP students had been tested. The real differences would have been larger because by definition the excluded students should be less able than the included students to perform on the test. To see how much larger the differences might have been, another analysis was performed with the assumption that

(Model VIII) the scores of excluded IEP/LEP students would have
been scored at the 0.01st percentile[9]--that is, only 1 out
of 10,000 students would have scored lower.

Under this assumption the average difference in mean scores for the jurisdictions would have changed by about 0.19 standard units (Table 2). The mean-score-based rankings for several states would have been changed considerably. Table 3 shows that the 12th ranked state would have moved up to 7th, the 7th would have fallen back to 10th, and the 10th to 14th. At the other end of the scale, the state ranked 3rd from the bottom (i.e., 36th) would have moved up to 8th from the bottom and the 9th from the bottom would have fallen to 3rd from the bottom (because testing in that state excluded relatively many IEP/LEP students compared to other states). The other changes in ranking were smaller, with the average change in rank equal to 1.7.[10] The percentages meeting the three performance standards would have changed by about 5 percent, 1 percent, and 0.0 percent respectively (Table 9). The rankings based on percentages of students meeting the performance standards change somewhat more at the basic level (average change in rank 2.6), about the same at the proficient level (1.65), and somewhat less at the advanced level (1.05); see Table 10. These results may be regarded as close to an upper bound for the differences in mean scores that actually would have occurred if all of the IEP/LEP students had been tested. One may conclude from these analyses that if it had been practicable to include all of the IEP/LEP students in the assessment, the states' scores and rankings would have changed moderately.

---

[9] The percentiles refer to the distribution of scores for the students eligible for testing in the TSA.

[10] The changes in rankings were as follows: 11 did not change, 10 changed one rank, 4 changed two ranks, 7 changed three ranks, 2 changed four ranks, 3 changed five ranks, and 1 changed six ranks.

Table 9. **Percentage of students scoring at or above basic, proficient, or advanced thresholds under alternative assumptions about IEP/LEP and private school students**

	Basic				Proficient				Advanced			
St.	obsvd	VII	VIII	IX	obsvd	VII	VIII	IX	obsvd	VII	VIII	IX
1	42.9	37.2	35.5	36.5	09.2	07.9	07.6	07.9	.2	.2	.2	.2
2	49.5	46.0	45.6	46.3	09.6	09.0	08.9	09.1	.1	.1	.1	.1
3	45.7	46.5	44.1	46.0	09.1	08.8	08.8	09.3	.6	.6	.6	.6
4	50.6	48.8	48.0	49.6	10.1	09.7	09.6	10.1	.3	.2	.2	.2
5	74.9	73.3	72.1	74.6	24.6	23.6	23.6	25.7	1.1	1.1	1.1	1.1
6	36.3	34.5	34.3	38.4	05.5	05.2	05.2	05.9	.2	.2	.2	.2
7	60.1	57.6	55.7	56.2	13.5	12.5	12.5	12.7	.3	.3	.3	.3
9	61.3	56.8	51.9	54.3	16.6	14.1	14.0	15.0	.8	.7	.7	.7
10	79.3	80.2	78.3	79.3	27.8	28.3	27.4	28.6	1.9	1.8	1.8	1.9
12	55.4	53.4	52.8	56.5	15.1	14.4	14.4	15.9	.8	.8	.8	.8
13	57.1	57.7	50.3	51.8	19.2	19.2	16.9	17.6	2.3	2.0	2.0	2.0
14	84.0	82.6	80.2	80.8	30.0	28.6	28.6	29.3	1.6	1.5	.5	1.5
15	55.4	54.9	52.4	55.5	17.1	17.4	16.2	17.6	1.3	1.3	1.3	1.3
16	76.7	70.0	63.1	65.1	25.6	22.0	21.0	22.3	1.8	1.5	1.5	1.5
17	53.0	42.8	39.3	40.7	12.5	09.3	09.2	09.7	.6	.4	.4	.4
18	73.7	72.1	71.5	71.9	19.2	19.3	18.7	19.0	.8	.8	.8	.8
19	60.4	61.4	57.3	59.7	15.9	15.1	15.1	16.2	.6	.5	.5	.5
20	58.2	57.5	56.2	59.3	14.9	14.8	14.4	15.8	.9	.9	.9	.9
21	49.5	49.6	49.6	52.0	13.9	13.9	14.0	15.0	.5	.5	.5	.5
22	55.6	53.5	52.8	54.0	12.7	12.0	12.0	12.5	.7	.6	.6	.6
23	73.8	71.2	71.0	72.2	21.5	20.8	20.8	21.7	1.2	1.2	1.2	1.2
24	66.3	65.7	64.6	65.6	18.4	17.9	17.9	18.5	.5	.5	.5	.5
25	57.1	45.1	41.7	45.5	16.4	12.0	12.0	13.3	1.3	1.0	1.0	1.0
26	43.8	39.4	38.7	42.8	12.3	10.9	10.9	12.2	.9	.8	.8	.8
27	65.8	62.9	62.3	64.0	17.9	16.9	16.9	17.9	1.1	1.1	1.1	1.1
28	69.7	64.6	62.3	63.3	23.2	21.1	20.7	21.3	1.2	1.1	1.1	1.1
29	63.9	62.9	62.3	65.5	19.1	19.4	18.6	20.4	.9	.9	.9	.9
30	66.3	64.5	64.4	67.1	22.7	22.3	22.1	23.9	1.8	1.7	1.7	1.7
31	67.2	66.3	65.2	67.5	23.1	22.5	22.3	24.0	1.8	1.7	1.7	1.7
32	73.8	66.9	63.7	64.0	20.4	17.7	17.6	17.7	.7	.6	.6	.6
33	77.0	76.1	76.0	77.2	24.6	24.3	24.2	25.5	1.8	1.8	1.8	1.8
34	53.1	50.9	50.8	52.2	15.0	14.4	14.3	14.9	.9	.8	.8	.8
35	50.9	49.0	44.2	46.3	10.4	09.0	09.0	09.6	.2	.2	.2	.2
36	84.4	76.6	67.8	68.8	31.3	26.2	25.0	25.7	2.1	1.7	1.7	1.7
37	53.5	49.3	48.4	52.2	16.0	15.0	14.6	16.1	1.1	1.0	1.0	1.0
38	19.4	18.4	18.4	21.9	02.6	02.5	02.5	02.9	.4	.4	.4	.4
39	50.7	45.8	44.2	44.9	10.3	08.9	08.9	09.1	.2	.2	.2	.2
40	48.0	44.9	42.5	45.0	12.7	11.7	11.2	12.0	.5	.5	.5	.5
AVERAGE DIFFERENCE FROM OBSERVED												
		2.9	4.6	3.4		1.1	1.3	1.0		.1	.1	.1

Table 10. **Rankings of states based on percentage of students scoring at or above basic, proficient, or advanced thresholds under alternative assumptions about IEP/LEP students and private school students**

	Basic				Proficient				Advanced			
St.	obsvd	VII	VIII	IX	obsvd	VII	VIII	IX	obsvd	VII	VIII	IX
1	36	0	0	1	35	1	1	1	37	0	0	0
2	32	-2	-4	-4	34	-2	0	1	38	0	0	0
3	34	-5	-3	-4	36	-1	-1	-3	25	-1	-1	-1
4	30	-2	-3	-3	33	-3	-3	-3	33	0	0	0
5	6	-1	-2	-2	6	-1	-1	-2	13	-1	-1	-2
6	37	0	0	-1	37	0	0	0	34	0	0	0
7	18	0	0	1	26	-1	-1	0	32	0	0	0
9	16	4	6	5	18	5	5	4	20	2	2	2
10	3	-1	-1	-1	3	-1	-1	-1	3	-1	-1	-1
12	24	-1	-5	-6	22	-1	-1	-2	21	-1	-1	-1
13	21	-4	3	5	13	0	3	4	1	0	0	0
14	2	-1	-1	-1	2	-1	-1	-1	8	0	0	0
15	23	-2	-2	-3	17	-1	0	-1	9	0	0	0
16	5	3	7	7	4	4	4	4	4	3	3	3
17	27	7	7	8	29	2	2	2	26	4	4	4
18	9	-3	-4	-3	12	0	-1	0	22	-1	-1	-1
19	17	-1	-1	-1	21	-3	-3	-3	27	-1	-1	-1
20	19	0	-2	-2	24	-4	-4	-3	17	0	0	0
21	31	-6	-6	-6	25	-1	-1	-2	29	-2	-2	-2
22	22	0	-2	0	27	-1	-1	0	24	-1	-1	-1
23	7	0	-1	-2	10	0	-1	-1	12	-2	-2	-2
24	12	-1	-3	-2	15	-1	-2	-2	30	-2	-2	-1
25	20	-12	-13	-11	19	8	8	6	10	5	5	5
26	35	0	0	-1	30	-1	-1	-2	19	0	0	0
27	14	1	-1	-1	16	1	-1	-2	14	-1	-1	-1
28	10	2	4	5	7	2	3	3	11	0	0	1
29	15	-1	0	-4	14	-3	-2	-3	16	0	0	0
30	13	0	-3	-4	9	-2	-2	-2	7	-2	-2	-2
31	11	-1	-3	-3	8	-2	-2	-2	6	-2	-2	-2
32	8	1	3	6	11	4	3	4	23	2	2	2
33	4	0	-1	-1	5	-1	-1	0	5	-2	-2	-2
34	26	-2	-3	-3	23	-1	-1	1	18	0	0	0
35	28	-1	1	1	31	2	1	1	36	0	0	0
36	1	2	6	6	1	2	2	2	2	4	4	4
37	25	1	1	-1	20	-1	-1	-1	15	-1	-1	-1
38	38	0	0	0	38	0	0	0	31	0	0	0
39	29	2	1	4	32	2	1	2	35	0	0	0
40	33	0	-1	-1	28	0	0	1	28	1	1	0
AVERAGE DIFFERENCE FROM OBSERVED												
		1.75	2.60	3.00		1.65	1.65	1.80		1.05	1.05	1.05

Separate analyses were also performed to provide insight into the scores and rankings if IEP students were tested and LEP students were not (Tables 11 and 12), and conversely if LEP students were tested and IEP students were not (Tables 13 and 14). The effect of including one group but not the other (IEP only or LEP only) tends to be about half the effect of including both IEP and LEP.

Combined Effect of Exclusion of Private School Students and IEP/LEP Students

An additional sensitivity analysis was run to estimate the differences in the mean standard scores and ranks of states that might have occurred if both private school students and the excluded IEP/LEP students had been tested. Testing IEP/LEP students would tend to lower scores, and testing private school students would raise them. Consider the effect on mean scores first. We will assume that

(Model IX) Model VIII and either Model II or Model V holds.

The average change in state mean standard scores in the simulation was observed to be 0.12, and the average change in ranks was 1.9, with some states' rankings changing by as much as seven or eight places. The state ranked second from the bottom advanced to 9th from the bottom and the 10th from the bottom fell to next to the last in rank. Changes at the top end were typically smaller, e.g., the 4th ranked state fell to 6th.[11] Turning now to percentages of students at the three performance levels, the changes were calculated to be 3 percent (basic), 1 percent (proficient), and 0.0 percent (advanced). By way of comparison, the standard errors (reflecting sampling variability) were 1 percent, 1 percent, and 0.1 percent respectively. The ranking of states according to percentage of students at the basic level changed substantially, by 3.0. The changes were somewhat less at the proficient level--1.8--but still appreciable, and they were smaller at the advanced level--1.05. *In all, these effects are much larger than the magnitude of sampling error* (Spencer 1991).

---

[11] The changes in rankings were as follows: 6 did not change, 12 changed one rank, 7 changed two ranks, 4 changed three ranks, 4 changed four ranks, 3 changed five ranks, 1 changed six ranks, and 1 changed seven ranks.

Table 11. **Percentage of students scoring at or above basic, proficient, or advanced thresholds under alternative assumptions about IEP and private school students**

	Basic				Proficient				Advanced			
St.	obsvd	VII	VIII	IX	obsvd	VII	VIII	IX	obsvd	VII	VIII	IX
1	42.9	38.3	38.0	39.0	09.2	08.3	08.2	08.5	.2	0.2	0.2	0.2
2	49.5	48.0	47.4	48.1	09.6	09.4	09.2	09.4	.1	0.1	0.1	0.1
3	45.7	45.8	44.7	46.6	09.1	09.1	09.0	09.5	.6	0.6	0.6	0.6
4	50.6	50.5	48.7	50.3	10.1	10.2	09.7	10.2	.3	0.3	0.3	0.3
5	74.9	75.0	73.6	76.0	24.6	24.6	24.5	26.6	1.1	1.1	1.1	1.2
6	36.3	35.9	35.5	39.6	05.5	05.5	05.4	06.1	.2	0.2	0.2	0.2
7	60.1	59.1	58.1	58.6	13.5	13.2	13.2	13.4	.3	0.3	0.3	0.3
9	61.3	57.5	56.2	58.4	16.6	14.8	14.7	15.8	.8	0.7	0.7	0.7
10	79.3	80.5	79.4	80.4	27.8	28.4	28.3	29.5	1.9	1.9	1.9	1.9
12	55.4	55.4	54.3	57.9	15.1	15.0	14.9	16.5	.8	0.8	0.8	0.8
13	57.1	57.1	55.2	56.6	19.2	19.2	18.6	19.4	2.3	2.3	2.3	2.3
14	84.0	83.9	83.0	83.5	30.0	29.9	29.9	30.6	1.6	1.6	1.6	1.6
15	55.4	56.2	55.1	58.1	17.1	17.4	17.3	18.8	1.3	1.4	1.4	1.4
16	76.7	71.0	69.9	71.6	25.6	22.8	22.7	24.0	1.8	1.6	1.6	1.6
17	53.0	54.3	53.2	54.6	12.5	12.9	12.8	13.3	.6	0.6	0.6	0.6
18	73.7	74.1	73.5	73.9	19.2	19.3	19.2	19.5	.8	0.8	0.8	0.8
19	60.4	59.9	58.6	60.9	15.9	15.7	15.6	16.8	.6	0.6	0.6	0.6
20	58.2	58.4	56.7	59.8	14.9	15.0	14.9	16.3	.9	0.9	0.9	0.9
21	49.5	50.9	49.7	52.1	13.9	14.3	14.1	15.1	.5	0.5	0.5	0.5
22	55.6	56.8	56.3	57.5	12.7	13.1	13.0	13.5	.7	0.7	0.7	0.7
23	73.8	74.0	72.1	73.2	21.5	21.6	21.4	22.3	1.2	1.2	1.2	1.2
24	66.3	66.9	65.5	66.4	18.4	18.7	18.5	19.1	.5	0.5	0.5	0.5
25	57.1	50.8	49.9	53.5	16.4	14.0	14.0	15.5	1.3	1.1	1.1	1.2
26	43.8	42.9	42.7	46.9	12.3	12.1	12.1	13.5	.9	0.9	0.9	0.9
27	65.8	65.4	63.5	65.2	17.9	17.6	17.6	18.6	1.1	1.1	1.1	1.1
28	69.7	66.9	66.1	67.1	23.2	21.7	21.6	22.3	1.2	1.2	1.2	1.2
29	63.9	64.5	62.7	65.9	19.1	19.2	18.5	20.4	.9	0.9	0.9	1.0
30	66.3	66.8	64.2	66.9	22.7	23.0	22.9	24.7	1.8	1.8	1.8	1.8
31	67.2	67.6	64.6	67.0	23.1	23.3	22.8	24.5	1.8	1.8	1.8	1.8
32	73.8	71.7	70.8	71.1	20.4	19.7	19.6	19.8	.7	0.7	0.7	0.7
33	77.0	78.1	76.9	78.1	24.6	25.1	25.0	26.2	1.8	1.9	1.9	1.9
34	53.1	53.0	52.3	53.7	15.0	14.9	14.7	15.3	.9	0.9	0.9	0.9
35	50.9	47.2	46.5	48.5	10.4	09.4	09.4	10.0	.2	0.2	0.2	0.2
36	84.4	78.9	77.5	78.2	31.3	27.1	27.0	27.8	2.1	1.8	1.8	1.8
37	53.5	51.3	50.9	54.6	16.0	15.3	15.2	16.9	1.1	1.0	1.0	1.1
38	19.4	18.4	18.4	21.9	02.6	02.5	02.5	02.9	.4	0.4	0.4	0.4
39	50.7	48.9	48.0	48.6	10.3	09.8	09.8	10.0	.2	0.2	0.2	0.2
40	48.0	46.4	44.9	47.4	12.7	12.2	11.9	12.8	.5	0.5	0.5	0.5
AVERAGE DIFFERENCE FROM OBSERVED												
		1.4	2.1	1.7		0.6	0.6	0.9		.0	.0	.0

Table 12. **Rankings of states based on percentage of students scoring at or above basic, proficient, or advanced thresholds under alternative assumptions about IEP students and private school students**

	Basic				Proficient				Advanced			
St.	obsvd	VII	VIII	IX	obsvd	VII	VIII	IX	obsvd	VII	VIII	IX
1	36	0	0	1	35	1	1	1	37	0	0	0
2	32	−1	−1	0	34	0	0	1	38	0	0	0
3	34	0	0	1	36	−1	−1	−2	25	0	0	0
4	30	−1	−1	−1	33	−2	−1	−2	33	0	0	0
5	6	−1	−1	−1	6	−1	−1	−2	13	0	0	−1
6	37	0	0	−1	37	0	0	0	34	0	2	2
7	18	−1	−1	0	26	0	0	2	32	0	0	0
9	16	3	4	3	18	5	4	4	20	2	2	2
10	3	−1	−1	−1	3	−1	−1	−1	3	−1	−1	−1
12	24	−1	−1	−3	22	−2	−2	−2	21	−1	−1	−1
13	21	−1	0	2	13	1	0	1	1	0	0	0
14	2	−1	−1	−1	2	−1	−1	−1	8	0	0	0
15	23	−1	−1	−3	17	0	0	−1	9	0	0	0
16	5	4	4	3	4	4	4	4	4	3	3	3
17	27	−3	−3	−2	29	−1	−1	0	26	0	0	0
18	9	−3	−3	−3	12	0	0	1	22	−1	−1	−1
19	17	−1	−1	−1	21	−3	−3	−2	27	0	0	0
20	19	−1	−1	−2	24	−3	−3	−3	17	0	0	0
21	31	−4	−3	−3	25	−1	−1	0	29	−1	−1	−1
22	22	−1	−3	0	27	0	0	−1	24	−1	−1	−1
23	7	0	0	0	10	0	0	−1	12	−2	−2	−2
24	12	−1	−1	1	15	0	−1	0	30	0	0	0
25	20	8	7	7	19	6	6	4	10	2	2	1
26	35	0	0	−1	30	0	−1	−3	19	0	0	0
27	14	0	0	1	16	0	0	1	14	0	0	0
28	10	2	0	0	7	2	2	3	11	0	0	2
29	15	0	0	−1	14	−1	1	−3	16	0	0	0
30	13	0	0	−1	9	−2	−3	−3	7	−2	−2	−2
31	11	−1	1	0	8	−2	−1	−1	6	−2	−2	−2
32	8	0	0	1	11	0	0	1	23	1	1	1
33	4	0	0	0	5	−1	−1	0	5	−2	−2	−2
34	26	−1	−1	0	23	−1	0	1	18	0	0	0
35	28	4	4	3	31	2	2	1	36	0	−1	−1
36	1	2	2	2	1	2	2	2	2	4	4	4
37	25	1	1	−1	20	−1	−1	−2	15	0	0	0
38	38	0	0	0	38	0	0	0	31	0	0	0
39	29	1	1	1	32	0	−1	1	35	0	−1	−1
40	33	0	0	0	28	1	2	2	28	1	1	1
				AVERAGE DIFFERENCE FROM OBSERVED								
		1.25	1.20	1.30		1.20	1.20	1.50		0.65	0.75	0.80

Table 13. **Percentage of students scoring at or above basic, proficient, or advanced thresholds under alternative assumptions about LEP and private school students**

	Basic				Proficient				Advanced			
St.	obsvd	VII	VIII	IX	obsvd	VII	VIII	IX	obsvd	VII	VIII	IX
1	42.9	43.9	43.6	44.6	09.2	09.3	09.3	09.5	.2	0.2	0.2	0.2
2	49.5	49.1	48.5	49.2	09.6	09.5	09.3	09.5	.1	0.1	0.1	0.1
3	45.7	47.6	46.4	48.4	09.1	09.0	08.9	09.4	.6	0.6	0.6	0.6
4	50.6	49.8	48.0	49.5	10.1	09.9	09.5	09.9	.3	0.3	0.3	0.3
5	74.9	75.0	73.6	75.9	24.6	24.5	24.4	26.6	1.1	1.1	1.1	1.2
6	36.3	35.5	35.1	39.2	05.5	05.4	05.3	06.0	.2	0.2	0.2	0.2
7	60.1	59.8	58.7	59.2	13.5	13.2	13.2	13.4	.3	0.3	0.3	0.3
9	61.3	61.9	60.5	62.6	16.6	16.6	16.6	17.7	.8	0.8	0.8	0.8
10	79.3	82.2	81.1	82.0	27.8	29.4	29.3	30.6	1.9	1.9	1.9	1.9
12	55.4	57.3	56.3	59.8	15.1	15.8	15.8	17.4	.8	0.8	0.8	0.9
13	57.1	59.1	57.2	58.6	19.2	19.8	19.2	20.0	2.3	2.1	2.1	2.1
14	84.0	84.8	84.0	84.4	30.0	29.9	29.9	30.6	1.6	1.6	1.6	1.6
15	55.4	56.3	55.2	58.2	17.1	18.0	17.8	19.4	1.3	1.3	1.3	1.4
16	76.7	78.2	77.1	78.5	25.6	26.1	26.0	27.4	1.8	1.9	1.9	1.9
17	53.0	44.4	43.4	44.9	12.5	09.7	09.5	09.9	.6	0.5	0.5	0.5
18	73.7	74.1	73.5	73.9	19.2	19.9	19.8	20.1	.8	0.8	0.8	0.8
19	60.4	62.9	61.6	63.8	15.9	15.7	15.6	16.7	.6	0.6	0.6	0.6
20	58.2	58.8	57.2	60.2	14.9	15.4	15.3	16.8	.9	0.9	0.9	0.9
21	49.5	51.3	50.1	52.6	13.9	14.5	14.3	15.3	.5	0.5	0.5	0.5
22	55.6	55.5	55.0	56.3	12.7	12.5	12.5	12.9	.7	0.7	0.7	0.7
23	73.8	74.0	72.1	73.2	21.5	22.2	21.9	22.9	1.2	1.3	1.3	1.3
24	66.3	67.3	65.9	66.9	18.4	18.5	18.4	19.0	.5	0.5	0.5	0.5
25	57.1	54.2	53.3	56.8	16.4	15.4	15.4	17.0	1.3	1.3	1.3	1.3
26	43.8	42.4	42.2	46.3	12.3	11.8	11.8	13.2	.9	0.8	0.8	0.9
27	65.8	64.0	62.1	63.8	17.9	17.6	17.6	18.6	1.1	1.1	1.1	1.1
28	69.7	70.2	69.4	70.3	23.2	23.8	23.7	24.4	1.2	1.3	1.3	1.3
29	63.9	64.2	62.5	65.7	19.1	19.7	19.1	21.0	.9	0.9	0.9	1.0
30	66.3	65.5	63.0	65.7	22.7	23.2	23.1	24.9	1.8	1.8	1.8	1.8
31	67.2	67.3	64.3	66.7	23.1	23.1	22.7	24.4	1.8	1.8	1.8	1.8
32	73.8	71.7	70.8	71.1	20.4	19.3	19.2	19.4	.7	0.7	0.7	0.7
33	77.0	77.6	76.4	77.6	24.6	25.2	25.0	26.3	1.8	1.9	1.9	1.9
34	53.1	52.2	51.5	52.9	15.0	14.7	14.4	15.0	.9	0.9	0.9	0.9
35	50.9	54.0	53.3	55.2	10.4	10.3	10.3	10.9	.2	0.2	0.2	0.2
36	84.4	84.5	83.1	83.7	31.3	31.9	31.7	32.6	2.1	2.1	2.1	2.1
37	53.5	54.1	53.6	57.3	16.0	16.6	16.5	18.3	1.1	1.1	1.1	1.1
38	19.4	19.5	19.5	23.1	02.6	02.6	02.6	03.1	.4	0.4	0.4	0.5
39	50.7	49.1	48.2	48.9	10.3	09.8	09.8	10.0	.2	0.2	0.2	0.2
40	48.0	47.7	46.2	48.7	12.7	12.5	12.2	13.1	.5	0.6	0.6	0.6
				AVERAGE DIFFERENCE FROM OBSERVED								
		1.2	1.5	1.8		0.5	0.5	1.0		.0	.0	.0

Table 14. **Rankings of states based on percentage of students scoring at or above basic, proficient, or advanced thresholds under alternative Assumptions about LEP students and private school students**

	Basic				Proficient				Advanced			
St.	obsvd	VII	VIII	IX	obsvd	VII	VIII	IX	obsvd	VII	VIII	IX
1	36	-1	-2	0	35	0	0	-1	37	-1	-1	-1
2	32	-2	-3	-2	34	0	0	1	38	0	0	0
3	34	-1	-2	-1	36	0	0	0	25	0	0	0
4	30	-1	1	-1	33	-2	0	0	33	0	0	0
5	6	0	0	0	6	0	0	-1	13	0	0	0
6	37	0	0	0	37	0	0	0	34	3	3	3
7	18	0	0	2	26	0	0	0	32	0	0	0
9	16	1	1	1	18	1	0	1	20	1	1	1
10	3	0	0	0	3	0	0	0	3	0	0	0
12	24	-3	-3	-5	22	-2	-2	-2	21	-2	-2	-2
13	21	-2	-2	0	13	-1	-1	0	1	1	1	1
14	2	-1	-1	-1	2	0	0	0	8	0	0	0
15	23	-1	-1	-1	17	-1	-1	-2	9	0	0	0
16	5	-1	-1	-1	4	0	0	0	4	0	0	0
17	27	7	8	8	29	4	3	3	26	4	4	4
18	9	-2	-2	-2	12	-1	-1	0	22	0	0	0
19	17	-1	-1	-2	21	0	0	2	27	-1	-1	-1
20	19	1	1	-1	24	-1	-1	-2	17	0	0	0
21	31	-3	-3	-3	25	0	0	-1	29	-1	-1	-1
22	22	1	1	3	27	1	0	2	24	0	0	-1
23	7	1	1	1	10	0	0	0	12	-2	-2	-2
24	12	-1	-1	-1	15	0	0	1	30	-1	-1	-1
25	20	4	5	4	19	3	3	2	10	1	1	1
26	35	1	1	-1	30	-1	-1	-3	19	1	1	1
27	14	1	1	2	16	1	1	1	14	0	0	1
28	10	0	0	0	7	0	0	1	11	1	1	1
29	15	-1	-1	-1	14	-1	0	-3	16	0	0	0
30	13	0	0	0	9	-1	-1	-2	7	0	0	0
31	11	1	1	1	8	1	1	1	6	0	0	0
32	8	1	1	1	11	3	2	3	23	0	0	1
33	4	1	1	1	5	0	0	1	5	0	0	0
34	26	1	1	1	23	1	1	2	18	0	0	0
35	28	-2	-2	-2	31	-1	-1	-1	36	-2	-2	-2
36	1	1	1	1	1	0	0	0	2	-1	-1	-1
37	25	0	-1	-2	20	-2	-1	-2	15	0	0	-1
38	38	0	0	0	38	0	0	0	31	0	0	0
39	29	2	1	2	32	0	-1	-1	35	0	0	0
40	33	-1	0	-1	28	-1	0	0	28	-1	-1	-1
AVERAGE DIFFERENCE FROM OBSERVED												
		1.20	1.30	1.40		0.75	0.55	1.05		0.60	0.60	0.70

Who Should Be Assessed?

A question that should be more closely addressed in the design of NAEP is: Who should the assessment cover? Currently, practical testing considerations (who is easy to test) determine who gets tested. The current procedure is not logically consistent. For example, a student with relatively low mental ability but who is not classified as mentally retarded will score low and bring down the state's mean. On the other hand, if the student's mental abilities are even lower (enough so that the student is classified as mentally retarded), the student will not be assessed and the state's mean will be higher (other things being equal). This does not make sense. Granted, there are practical difficulties in assessing IEP/LEP students but to simply exclude them from the population to whom inferences are being made seems inconsistent.

The issue is not purely a statistical matter but involves how the NAEP data are used for drawing inferences and making decisions about the nation's students. If the inferences and decisions would be better served by testing the IEP/LEP students, then consideration should be directed to developing assessment procedures that can cost-effectively estimate their test performance. For example, the IEP/LEP students who are difficult to assess could be sampled at much lower rates than other students. Thus, the assessment of such IEP/LEP students need not be unduly expensive and could involve less data collection and more statistical modeling than is currently done.[12] If the assessments are to continue to exclude IEP/LEP students then better criteria for exclusion need to be developed. "Better" criteria means criteria that are readily understood by those who interpret the data and base actions on them. "Better" criteria would also be amenable to consistent application within and across states. *If future assessments will exclude the IEP/LEP students, then they should provide for analysis of the excluded students in the sample. The analysis ought to compare the characteristics of excluded and nonexcluded students, so that persons analyzing the assessment data will understand which kinds of students are being excluded. The analysis also ought to reassess the validity of the exclusion criteria and to investigate the reliability with which the criteria are applied.*

---

[12] The statistical models that seem appropriate involve methods for "small area analysis," in which the sample sizes for IEP/LEP students in individual states are relatively small.

As a matter of statistical policy, the choice of the population to be tested should be based on the (a) information needs of the educational policymakers, decisionmakers, and others (b) analysis of how the data will be used. Once those two analyses are conducted, the statistical task is to estimate the test scores as accurately as possible, given the inevitable cost constraints. If the information needs are such that the relevant population is difficult to assess, then so be it. For the same cost of NAEP, more accurate statistics can be produced if the right population is surveyed than if the population to be surveyed is restricted for data collection considerations. The sampling errors of the statistics may be higher because the testing is more difficult but the total error will be lower (if the survey is properly designed).

For these reasons *attention should be given to changing the scope of NAEP to encompass IEP/LEP students.* The current strategy seems based on an overly simplistic "cost-benefit analysis." That is, since it would be relatively expensive to assess IEP/LEP students, they should be excluded from NAEP. A more reasonable and thoughtful cost-benefit analysis should be conducted.

Summary and Recommendations

The design of the TSA allowed for the exclusions of three groups of students: students enrolled in private schools, students with limited English proficiency (LEP) who were judged incapable of taking the test, and special education students with individualized education plans (IEP) who were judged incapable of taking the test. Nationwide, *one in six* students belongs to one or more of these groups. States vary widely in the proportions of students belonging to these groups. Sensitivity analyses indicate that the exclusion of these groups of students could have important effects on state-by-state and state-to-national NAEP comparisons. Why should these groups be excluded from future state NAEP? There is no apparent justification for excluding private school students. Therefore *it is important that private school students be included in future state NAEP assessments.*

If, as seems likely, the decision to exclude IEP/LEP students is based on difficulty of assessing them, then the cost-benefit analysis implicit in such a decision needs to be reconsidered. The way that NAEP data are currently interpreted by almost all users implies that the excluded IEP/LEP students are assumed to have the same mean scores as the assessed students. Most users of the data would not explicitly make such an assumption and may object to it once it is exposed. However, the typical user of NAEP data is not aware of such "subtle" technical aspects of the data collection and acts as though the data pertained to the entire subpopulation of students (by grade, type of school, etc.). There are a variety of cost-effective methods of representing the IEP/LEP students in the assessment. The methods are based on lower sampling rates for expensively assessed students. Although the sample-based estimates for any single state will be less precise than desired (due to smaller sample sizes), the sample-based estimates can be adjusted in various ways to greatly increase their precision at the cost of a small amount of bias. The net result is increased accuracy. The adjustments can be achieved by using so-called "empirical Bayes" methods to "smooth" the observed state-to-state variation, (e.g., for IEP/LEP students now judged incapable of taking the test). The adjustments can also use external test data, where available. These techniques need not involve great additional expense and they would lead to considerably more accurate statistics. Therefore *consideration should be given to including all IEP/LEP students in the tested population.* If future assessments continue to exclude IEP/LEP students, then *more study should be given to the formulation of consistent and valid criteria for exclusion and to the implementation of the criteria.*

References

Office of Bilingual and Minority Affairs. (Undated). *A Summary of State Reports on the Number and Condition of Limited English Proficient Students, 1986-87.* Washington, DC: U.S. Department of Education.

Office of Special Education Programs. 1990. *Twelfth Annual Report to Congress.* Washington, DC: U.S. Department of Education.

Spencer, B. D. 1991. "A Critique of Sampling in the 1990 Trial State Assessment." (in this volume)

Spencer, B. D., Frankel, M. R., Ingels, S. J., Rasinski, K. A., and Tourangeau, R. 1990. *National Education Longitudinal Study of 1988 Base Year Sample Design Report.* (Data Series DR-NELS:88 - 88 - 1.7). Washington, DC: National Center for Education Statistics.

Westat, Inc. 1990. *Summary for All States of Training and Data Collection Activities, Preliminary Results, 1990 Trial State Assessment.* Rockville, MD.

Implementation of the 1990 Trial State Mathematics Assessment

Donald H. McLaughlin
Frances B. Stancavage
Jay G. Chambers
Elizabeth Hartka
Kadriye Ercikan
American Institutes for Research

August 1991

Executive Summary

In support of the National Academy of Education's evaluation of the Trial State Assessment (TSA), the American Institutes for Research (AIR) carried out a study of the administration of the 1990 TSA. The purpose of this study was to assess the feasibility of the full implementation in terms of state and local costs and the impact of state and local variations in implementation of the assessment procedures on the validity of the results. The aims of the study are both to provide an independent assessment of potential threats to the credibility of the state-level data and to identify specific areas for the refinement of the TSA in the future. The 1990 TSA was carefully planned and implemented, with substantial training of local administrators. Nevertheless, refinement of the TSA implementation should be considered in several areas in 1992.

AIR used a variety of data sources for the study, including: (1) performance data on the TSA sample of 100,843 assessed students, as well as on a matched NAEP comparison sample of 2,487 students generated by the NAEP contractor, Educational Testing Service (ETS); (2) quality control monitor (QCM) data; (3) excluded student data on 5,735 IEP and LEP students who did not take the test (many IEP and LEP students did take the tests); (4) school questionnaire data and school substitution information; (5) state coordinator survey data; and (6) a sample of administration schedules (without student names), all provided by ETS, Westat, and National Computing Systems. Additional sources included (7) an AIR survey of state testing directors and (8) notes from AIR observations of TSA training and administration sessions. To assess validity, AIR computed both frequencies of "irregularities" and comparisons of performance related to these "irregularities." (For the purpose of this study, "irregularities" refer to any variations which were recorded; no presumption is being made as to the relative importance of particular irregularities.)

In making performance comparisons, AIR focussed both on mean levels and levels of variance, because it is as important to identify correlates of increases in variance, within session, between schools, and between states, as it is to identify correlates of mean difference. For example, a finding of greater variation between schools within states among unmonitored, as opposed to monitored sessions, would call for greater emphasis in LA training on uniformity in those areas in which the presence of monitors reduces variation.

In a study such as this, findings relating to session "irregularities" need not be computed so as to be nationally representative. Instead, it is more important to maximize the effective sample size for the comparisons. Although the number of students in a session was taken into account in computing statistical estimates, differential student weights were not employed. Thus, all sessions with the same number of students were given equal weight in the analyses. Furthermore, to eliminate sources of performance variance associated with race and sex, AIR carried out comparisons both on "raw" scores (actually, log(p/(1-p)), equated for booklet differences) provided by ETS and on "adjusted" scores from which race by sex means were subtracted. Typically, a difference of between .01 and .02 in a state mean on these scores would be associated with a unit change in state rank.

Overall Validity Estimation

First, a comparison was made between monitored and unmonitored sessions; and only very small differences were found, in either mean levels or variances, using either raw or adjusted scores. This is important because most of what is known about variations in session administration and their impact on performance is based on monitored sessions only. Second, a comparison was made with the NAEP comparison sample generated by ETS to match the overall TSA Sample. Because the comparison sample over-represented minorities, it was essential to use the adjusted scores in this comparison. The aggregate average TSA performance was .12 standard deviations higher than the aggregate average comparison sample performance, on a measure with a nominal population standard deviation of 1.00. This means that the aggregate average based on the matched comparison sample was higher than 10 states and lower than 30 states in the TSA sample.

Participation Rates

Approximately 6 percent of all originally invited schools refused to participate in the TSA, although in two states the proportion approached 20 percent. Schools that refused tended (1) to be smaller and (2) to have fewer black students than other schools. Westat successfully matched substitute schools to refusing schools on demographic factors. Performance scores in substitute schools were only .02 standard deviations higher than in cooperating original schools (.10 higher without the race, sex adjustment, however). This level of difference would not be expected to account for any significant change in state ranks.

On the other hand, initial examinations of data on the 5,735 students explicitly excluded because of severe disabilities or English language deficiencies revealed no substantial factors that might account for performance variations. Numbers of these IEP and LEP students excluded in individual states ranged from 48 (4 percent) to 244 (9 percent) across the 40 states. Of 3,915 sessions, only 119 had more than 5 excluded students and of these, only 22 had more than 10.

In addition to explicitly excluded students, there were students who were no longer at the school (withdrawn) or absent (and who did not have a make-up session). It is impossible to know how withdrawn or absent students would have performed, but evidence was found relating numbers of withdrawn or absent students to scores of those present for the assessment. There was a significant negative correlation, -0.27, between the number of students withdrawn from a session and the average performance score of those present in that session, and virtually the same negative correlation was observed between numbers of absent students and average scores. The most straightforward interpretation of these results is that schools with high rates of student absences and departures are also schools with lower mathematics achieving students.

It should be noted that one state differed from all others in the frequency of absences, and the presentation of results for that state should be qualified by the large absence rate.

Pre-Assessment Variations

Quality control monitors (QCMs) made pre-assessment check-up calls to all local session administrators a few days prior to the assessment, to ensure receipt of materials and completion of preliminary steps, such as distribution of teacher questionnaires. In the majority of cases, everything was being carried out in accordance with plans at the time of the call. Furthermore, performance scores were not significantly related to deviations from these procedures. Nevertheless, these calls are important because LAs who needed the reminder to complete pre-assessment procedures also tended to require assistance during the session. Finally, in many cases the teacher questionnaires had not been distributed, and the return rate for this rich source of contextual information would undoubtedly have been lower had the reminder not been given.

Session Administration

Irregularities in session administration affect validity primarily by creating especially favorable, or especially unfavorable testing conditions for particular students. For example, inaccuracies in the timing of the subtests could result in some students having more or less time than intended to complete certain testing tasks, and unfair opportunities to copy from the work of others might arise as a result of inadequate supervision by testing monitors and/or crowded seating. Deviations by the LAs in reading instructions might create confusion or misunderstanding among some students so that they could not properly execute the testing tasks. On the other hand, unsanctioned elaborations on instructions, or pep talks, might provide an unfair advantage.

For most of these potential problems, information on the frequency of occurrence is available on the forms filled out by the QCMs. However, since the QCMs were instructed to intervene to correct observed problems wherever possible, there is only limited opportunity to assess the magnitude of the impact on student scores in the absence of such intervention. Specifically, in addition to an overall judgment, the QCM forms provide information on several aspects of LA performance that are pertinent to session administration, such as reading instructions, timing sections, instruction on use of calculators, handling students' questions, and monitoring students' work.

In addition, information is available on a few conditions of the administration which may or may not be under the control of the LA, but which also have the potential to affect student performance, such as time of day, adequacy of the room, and attitude and behavior of the students participating in the assessment.

Among those factors under the control of the LA, the most frequent irregularities were errors in reading the test instructions (additions, omissions, or revisions) and errors in distributing and collecting the calculators between test sections. Errors in three or more script sections (out of five) were observed in 13 percent of the monitored assessment sessions, and errors in calculator distribution were observed in 11 percent of the sessions. Sessions with errors in three or more script sections had significantly lower student performance, with a mean adjusted z score of -0.084 compared to 0.024 for the remainder ($t = 3.42, p \leq .001$). No decrement in student performance was associated with errors in calculator distribution, but these results might have been different if the QCMs had not intervened in virtually all instances of distribution error.

After errors in script reading, the single factor most clearly associated with lower student performance was student disruption. Among the 7 percent of sessions where the QCM indicated the presence of student disruption, the mean adjusted z score was -0.091 compared to 0.018 for the remainder ($t = 3.58$, $p \leq .0004$). Fortunately for the integrity of the state-by-state comparisons, these and other session administration irregularities tended to be distributed randomly over the participating states, with no state as a significant outlier.

Incentives

In 20 percent to 25 percent of the monitored sessions, administrators met with students prior to the assessment. In 15 percent of the cases they gave a pep talk, in 9 percent of the cases they gave students an incentive to participate, and in 5 percent of the cases they gave special instructions in using calculators. Generally, students who were given incentives, a pep talk, or special instructions performed more *poorly* than others--a reminder of the difficulty of making causal inferences from data such as these. Examination of the effects of incentives is continuing.

Sessions were conducted by counselors, teachers, principals, and other educators. Scores in sessions conducted by principals, vice principals, and superintendents were higher than in other sessions. The most straightforward interpretation of this finding is that the apparent position of power of the test administrator may have affected students' perceptions of the importance of good performance on the test.

State and Local Costs for TSA

Rough estimates of the amount of effort involved in the various tasks required for participation in the TSA were gathered by AIR. Most state testing directors perceived their costs generally to be in line with their expectations; however, retrospective cost estimates varied widely among states because of the many ways in which components of the cost could be inadvertently overlooked (e.g.,the time spent by mathematics teachers to fill out questionnaires). An initial analysis of costs by function suggested that the total costs for a state's participation were probably between $90,000 and $210,000, most of which were incurred at the local level. Most state testing directors felt that the assessment was worth the cost; and the number of participating states is expected to increase in 1992.

Overall, the initial evaluation of the TSA indicates that the state NAEP is probably feasible and probably will provide the basis for reporting valid state-by-state results. A number of issues and concerns were raised, however, by the patterns observed in the data, and these should be addressed. First, monitoring of sessions provided valuable data and probably avoided many problems, and it should be continued. Second, issues such as the differences between the Trial State Assessment scores and the national NAEP scores need further investigation, especially as they relate to student motivation during the assessment. Finally, training of local administrators (LAs) should emphasize, even more, the value of uniformity of administration procedures. Although the administration was of high quality, the infrequent variations that did occur were often associated with lower average performance scores.

Introduction

Since 1968, the National Assessment of Educational Progress (NAEP) has provided a continuing national profile of the achievement of children and youths in American schools. A similar profile of educational progress within and among the states has not been available until recently, however. In April 1988, Congress authorized the expansion of NAEP to include a Trial State Assessment (TSA) of eighth grade mathematics during 1990. The trial is a voluntary assessment program administered by participating states under the direction of the U.S. Department of Education through the NAEP contractor.

Using the same Assessment instruments employed in the 1990 National Assessment of Educational Progress in Mathematics, the Department worked jointly with 40 state education agencies (37 states, 2 territories, and the District of Columbia) to select schools and students to participate in the 1990 TSA. State directors of testing selected local school personnel to be trained in NAEP testing methods and to administer the tests.

In each participating state the 1990 Trial State Assessment involved:

- a testing period from February 5 and March 2, 1990
- a sample of about 100 public schools that had an eighth grade (In states that had fewer than 100 schools with an eighth grade, all eighth grade schools were asked to participate.)
- assessment of 25 to 30 sampled students per school (In schools with fewer than 35 eighth grade students, all eighth graders were asked to participate.)
- about one hour of each sampled student's time
- completion of Excluded Student Questionnaire for each sampled student who was excluded from the assessment

- completion of Teacher Questionnaire by each eighth grade mathematics teacher with students in the assessment sample

- completion of School Characteristics and Policies Questionnaire and a Principal Questionnaire by the school's principal

The legislation authorizing the trial assessments requires that they be evaluated to determine "the feasibility and validity of [state] assessments and the fairness and accuracy of the data they produce." Ensuring that the assessment is administered under reasonable and uniform conditions for all students is an important factor in the fairness and accuracy of the resulting data. If the assessment can be shown to provide bona fide state-by-state estimates of academic progress, and if the data collection methods can be shown to be feasible, then a vehicle exists for generating important state-level educational productivity indicators. Careful examination of the 1990 TSA will allow the Department and state education agencies to address many questions about the validity and feasibility of state assessments. It will also provide an opportunity to identify additional data needs for the 1992 TSA. Thus, this report includes a summary of findings from the 1990 TSA, as well as a discussion of questions that cannot be answered with available data.

This report addresses whether and to what extent variations in implementation of 1990 TSA procedures might have affected the quality of the data obtained from the assessment. However, the report does not address either psychometric or statistical estimation issues. Furthermore, it does not address issues relating either to the determination of the content of the assessment or to the manner of reporting outcomes of a state assessment. These issues will be addressed in other reports from the TSA Evaluation Panel. The intention of the present report is to provide useful information for decisions concerning future NAEP state assessments, in terms of both refinements of procedures and collection of supplementary quality control data.

The U.S. Department of Education was responsible for preparing the assessment materials for the 1990 TSA, for selecting the sample of schools in each participating state, for providing training and written instructions for procedures to be carried out under the authority of state education agencies, and for compiling the results of the assessment. The Department had previously selected the Educational Testing Service (ETS) as a contractor (with Westat, Inc., serving as subcontractor) to conduct the national assessment; these agencies played major roles in the conduct of the TSA.

In October 1989, each participating state designated a state coordinator to manage the assessment within the state, and a Westat staff member, designated as state supervisor, was assigned to work with the state coordinator to implement the assessment. The state coordinator, in turn, obtained the commitment of selected schools to participate and named a local administrator (LA) to conduct the assessment in each participating school. The state coordinator (usually the state testing director) obtained a list of eighth graders from each participating school, and the state supervisor selected a sample to be tested from this list, using explicit formal sampling procedures. The lists of sampled students were sent to the LAs, who in turn notified the students, obtained parental consent forms as needed, and set up the testing sites and times.

The state supervisor, assisted by other contractor staff, trained the LAs in the assessment procedures in a day-long session that covered pre-assessment activities, conduct of the assessment, and post-assessment activities. The training materials included a videotape explaining the importance of NAEP and its assessment principles, as well as explicit exercises to be completed by LAs and reviewed by trainers.

LAs received and assumed responsibility for the security of the assessment materials shipped by the NAEP contractor and distributed questionnaires to school officials and mathematics teachers. The LA stored the test booklets, unopened and sealed, until 45 minutes before the beginning of the test session. Meanwhile, the LA examined the list of selected students to determine which, if any, should be excluded because they could not participate meaningfully in the assessment. The LA selected a teacher familiar with an excluded student to complete a questionnaire about that student. Finally, the LA prepared a room for the testing session.

A contractor's quality control monitor (QCM) checked on progress of preparations for the testing a few days before the session and obtained instructions from the LA on how to get to the testing room. LAs did not know until an hour before the session whether or not the QCM would observe the testing. If the state supervisor or the QCM judged that there was some likelihood that the LA would be unable to conduct the session in accordance with the guidelines, provision was made for the QCM to assist in the administration.

Data Sources

The report on the administration of the 1990 TSA is based on (1) direct observation of assessment sessions; (2) the analysis of data provided on tape by ETS, the NAEP prime contractor, and its subcontractor, Westat; (3) the analysis of data collected in an independent survey of state testing directors; and (4) the examination of a sample of paper records, such as administration schedules and quality control monitor debriefings.

The data provided by ETS consisted of six files: a student file, school file, and excluded student file for the TSA; and three corresponding files for a matched sample from the concurrent national assessment. The matched sample was selected by ETS. Although weights employed for estimating means and variances were computed by ETS, comparisons in this report are based on unweighted data. The reason is that the purpose of this report is to compare sessions of different types (e.g., monitored and unmonitored). Variations in administration of the assessment are of equal importance for the purpose of identifying problems in the TSA, whether they are in sessions with large or small weights. In the case of comparisons between the TSA data and the comparison sample of the NAEP data, the decision to carry out comparisons unweighted required eliminating the effects of the oversampling of racial-ethnic minorities on performance measures in the NAEP data. This was accomplished by subtracting race, as well as sex, means from the performance measure used prior to analysis. Computational formulas for variance components and significance tests are shown in Appendix A.

In addition, hard copy data and computer file data were provided to AIR by Westat on the quality control monitor questionnaires for both monitored and unmonitored sessions; the characteristics of substitute schools, as compared to the schools for which they were substituted; state supervisor's questionnaires; numbers of sampled, withdrawn, absent, makeup, and assessed students in each session; and a sample of administration schedules.

Data were also collected by AIR in an independent survey of state testing directors. The types of data collected are shown in Appendix B. A mail survey of all 50 states, the District of Columbia, and 4 territories during July 1990 obtained valid responses from all but one (non-participating) state. A follow-up survey containing a limited set of cost questions was conducted with a subsample of states during December 1990.

Results

The evaluation addressed two categories of issue: validity and feasibility. The issues concerning validity were addressed primarily through comparisons between monitored and unmonitored sessions; between TSA results and results of the matched NAEP sample; between related items that indicated discrepancies; and between written records and information on tape. Issues concerning cost were addressed primarily through analysis of questionnaires concerning the amount of time taken for tasks involved in TSA. These included the AIR survey as well as the Westat survey of state coordinators and debriefings of QCMs and LAs.

Validity Evaluation

Providing each state with an equal opportunity to measure the capabilities of its students requires that a representative sample of students be tested in each state and that these students be tested in an identical manner. As part of the overall TSA design, procedures and materials were naturally developed to ensure, to the greatest extent possible, that these conditions would be met. In evaluating the validity of the results and planning for future administrations, one needs to consider the nature and frequency of procedural irregularities that occurred despite these precautions, as well as the actual impact of specific types of irregularities on student performance. Moreover, since the results are for use at the state level, it is also essential to determine whether irregularities, if observed, were evenly distributed across states or whether they tended to occur with significantly greater frequency in some states. Finally, one needs to consider how the existing procedures and materials could be enhanced to further safeguard against the most damaging irregularities, as well as to determine where and how procedures might be simplified (and costs reduced) without compromising validity.

A major concern for the TSA is whether comparisons from state-to-state are credible: are the data collected in each state comparable to those collected in other states? In the past, NAEP has used a methodology that requires all data to be collected by a single contractor, and this methodology helps to ensure the uniformity and comparability needed for valid national estimates. In adapting NAEP to a state-by-state assessment, a methodology is being tested which (1) relies on compilation of student lists by state administrators, following uniform written instructions; and (2) involves collection of data by LAs, who are in turn required to undergo uniform training by a single contractor. The credibility of the state-by-state assessment hinges on whether variation in outcomes can be attributed to educational processes and student population differences, as opposed to differences in the implementation of the assessment.

There are three methods for assessing the overall validity of TSA data. First, one can compare the results of the TSA with the results of the concurrent NAEP. If the amount of between-state variation or between-school variation is significantly greater in the TSA than in NAEP, taking measurable differences into account, then there is cause for concern about the uniformity of TSA procedures. Second, quality control monitors (QCMs) observed one half of the assessment sessions. If either there were differences in mean levels of results between monitored sessions and unmonitored sessions or there was greater variation among unmonitored sessions than among monitored sessions, then there is concern that failing to monitor sessions may affect outcomes. Third, the content of debriefings of QCMs and of LAs can be used to identify specific non-uniformities. That debriefing information is particularly useful in attempting to explain unexpected variations and in identifying strategies for removing such variations, if they exist.

As an initial, overall comparison between TSA and NAEP, AIR compared scores on the TSA file provided by ETS with scores provided for the comparison sample, also selected by ETS. The results, for 3,915 TSA sessions and 108 comparison sample sessions, are presented in Table 1.

Table 1. Means and Variances of Performance Scores for the Trial State Assessment and a Matched National NAEP Sample (3915 TSA Sessions and 108 Comparison Sessions)

Standardized Logit Scores	Not Adjusted for Ethnic and Sex Effects	Adjusted for Ethnic and Sex Effects
Mean for TSA	-.000	-.000
Mean for Comparison Sample	-.180	-.118
Within-Session Variance for TSA	.802	.767
Within-Session Variance for Comparison	.936	.874
Between-Session Variance for TSA	.140	.078
Between-Session Variance for Comparison	.197	.106
Between-State Variance for TSA	.040	.032
Between-State Variance for Comparison	-[1]	.018
t-value for mean difference	2.59*	2.36*

[1] Insufficient data to estimate (negative estimate)

* Significant at $p \leq .05$

Several points must be understood to interpret this table correctly. First, the means for the TSA sample were forced to equal zero, within round-off error, as a matter of definition. The "unadjusted score" for each student in the TSA sample was defined in terms of p, the proportion of items that that student answered correctly. Specifically, it was $\log(p/(1-p))$, minus the mean of $\log(p/(1-p))$ for all students with the same booklet of items, divided by the standard deviation. Thus, the unadjusted mean and variance should be 0 and 1, respectively. The "adjusted" score further subtracted the mean of the unadjusted scores for all members of the same racial-ethnic and sex group in the TSA sample, ensuring that mean session performance, as measured by the adjusted score, was independent of the racial-ethnic and sex distribution of students in the session.

For comparability, the $\log(p/(1-p))$ scores for the comparison sample were also adjusted for the same means. The variance component computations and the Student's t tests of significance were based on formulas, shown in Appendix A, that take into account variations in session size, between sessions, and in numbers of sessions, between states.

If the samples were equivalent, then the distributions of the adjusted scores should be approximately the same, although the "unadjusted" scores for the comparison sample would be affected by the oversampling of ethnic minorities in the national NAEP. There are noticeable differences between the TSA sample and the comparison sample in Table 1, however. In particular, the difference in mean performance scores must be explained. Whether it is a result of differences in test administration procedures or of differences in the sample has important implications for the refinement of the Trial State Assessment.

To decide between these competing explanations, differences between the samples can be explored. Although the comparison is unweighted, the differences in adjusted means could not have been due directly to differences in race distributions between the TSA sample and the comparison sample. (The comparison was repeated using the weights provided by the NAEP contractor and the differences persisted.) Further, as a test of the extent to which the samples consisted of comparable students, the weighted means on "mother's education" were also compared. The two samples were virtually exactly matched on this measure, indicating that the differences between the two samples were directly related to performance in the assessment.

How large a difference is important for the TSA? The issue is not one of statistical significance because differences that are statistically significant on a sample of over 100,000 may still be of little interest. To determine the critical level of difference for these data, the distribution of (encrypted) state means was generated. In the middle two thirds of the distribution, a difference of roughly .05 in an unadjusted mean, or of .04 in an adjusted mean, would be sufficient to change a state rank by three. At the extremes of the distribution, a difference of roughly .10 (unadjusted) or .05 (adjusted) would be sufficient to change a state rank by one. In fact, about 30 of the 40 states were above the comparison sample mean and 10 were below. If this difference is not an artifact of the comparison, then there is a sufficient difference to warrant further investigation as to the cause.

In addition to mean performance levels, estimated components of variation in scores are also presented in Table 1. As with the means, these estimates are based on unweighted data, although differences in session size are taken into account, as shown in Appendix A. It is clear that *most* score variation is between students in a session, while the *least* score variation is between states. Components of variation of adjusted scores are smaller, as expected, with a very noticeable decline in between-state variation. From this, it is possible to conclude that *a substantial proportion of between-state score variation is associated with different racial-ethnicity distributions.* Finally, within-state variation in scores in the comparison sample is greater than variation in the TSA sample.

The second overall comparison of interest was between monitored and unmonitored sessions. This comparison is especially important for deciding the amount of monitoring required in the future. Should substantial score variation occur between monitored and unmonitored sessions, then reducing monitoring effort is not warranted. The results, for 1,918 monitored sessions and 1,997 unmonitored sessions, are shown in Table 2.

Table 2. Means and Variances of Performance Scores for Monitored and Unmonitored Sessions

Standardized Logit Scores	Not Adjusted for Ethnic and Sex Effects	Adjusted for Ethnic and Sex Effects
Mean for Monitored Sessions	.009	.010
Mean for Unmonitored Sessions	-.009	-.010
Within-Session Variance for Monitored Sessions	.803	.766
Within-Session Variance for Unmonitored Sessions	.800	.768
Between-Session Variance for Monitored Sessions	.137	.076
Between-Session Variance for Unmonitored Sessions	.144	.080
Between-State Variance for Monitored Sessions	.087	.031
Between-State Variance for Unmonitored Sessions	.091	.033
t-value for mean difference	1.24	1.86

The differences are very small. Finding slightly lower mean levels in unmonitored sessions may be expected because (1) generally, inadvertent errors in administration tend to reduce student performance, and (2) monitors, who were themselves trained administrators, were instructed to intervene if they perceived any variation that would affect performance. Finding slightly smaller variances in monitored sessions is consistent with the hypothesis that the administrations were probably more uniform in monitored sessions.

The finding of only very small differences between monitored and unmonitored sessions is crucial for the interpretation of irregularities observed and other factors noted by the QCMs in the sessions that they monitored. Generalizing findings from the monitored sessions to the entire assessment, it is necessary to make the assumption that the conduct of monitored and unmonitored sessions was essentially the same.

We turn now to a detailed analysis of validity of the TSA, based primarily on the reports of the QCMs. Non-uniformity of procedures can affect the outcomes of the assessment in four ways: (1) they may affect who takes the assessment (selection factors); (2) they may affect participating students' preparation for the assessment (incentive factors); (3) they may affect the interaction of the students with the assessment materials (session factors); and (4) they may affect the inclusion of student performance data in the resulting files (transmission factors).

Validity Issues Related to Sampling and Participation

The credibility of TSA comparisons between states crucially depends on uniformity of representation of each state's eighth grade population in the assessed sample. In a separate report of the Panel, the validity of the sampling plan is evaluated.[1] In the present report, issues related to the practical implementation of the sampling procedures and its relation to outcomes of the assessment are addressed. The major issue is whether types of students who might be expected to score lower on the average were less likely to be included in the assessment in some states than in others. For example, there is ample evidence that non-Hispanic Asian and white males score higher on math tests and although the fact that states differ in proportions of these populations does not invalidate comparisons, significant deviations between the assessed sample and the state population in a state would threaten the validity of interpretation of that state's scores. (A separate issue concerning the reporting and interpretation of TSA results is whether states should be compared separately for populations of "similarly situated" students--e.g., students with the same racial-ethnic and sex distributions.)

School participation rates. The state coordinators were requested to obtain full (100 percent) cooperation from the sampled schools. However, if some schools refused, contractor staff were prepared to select a few replacement schools. The "after-substitution" figure reflects the participation rate after the replacement schools were included in the sample. Substitutes were chosen on the basis of their similarity to the original schools on the variables used for stratification, and their eighth grade enrollment. Overall, 94 percent of sampled schools agreed to participate. Replacements were selected for the remainder, and many of them participated, yielding a final school participation rate of 97 percent.

---

[1] Spencer, B.D. *A Critique of Sampling in the 1990 Trial State Assessment*, in this volume.

The frequency distribution of school substitutions is shown in Table 3. In most states, no substitutions were required, and in only two states were there a large number of substitutions (23 out of 112 and 19 out of 107).

Table 3. State Frequency Distribution of School Substitutions

Number of Substitutions	Number of States
0	26
1	1
2	2
3	1
4	2
6	1
8	2
9	2
11	1
19	1
23	1

In every case but one, a replacement school in the same state was selected for each refusing school. The questions arise as to how different the refusing schools in each state were from those that did not refuse and how different the replacement schools in each state were from the refusing schools. These comparisons are displayed in Tables 4 and 5. For example, according to the first row of Table 4, in 11 states the mean eighth grade enrollment was greater in cooperating schools than in refusing schools, while in 3 states the reverse was true. In the 5 other states in which some schools refused, the mean was the same.

Table 4. **State Frequency Distributions of Originally Cooperating and Refusing Schools (For the 19 states where 1 or more refusals occurred)**

	Greater in Cooperating Schools	Greater in Refusing Schools
Mean Eighth Grade Enrollment	11	3
Mean Percentage of black Students	12	2
Mean Percentage of Hispanic Students	7	7
Median Income	7	7

The differences between cooperating and refusing schools were not under the control of the sampling system; on the other hand, substitute schools could be matched to the refusing schools on demographic factors. The main differences between cooperating and refusing schools, which were statistically significant in some states, were that smaller schools tended to refuse more frequently, and that schools with lower percentages of black students tended to refuse more frequently. As shown in Table 5, substitute schools were selected to match the refusing schools. No differences within a state were statistically significant.

Table 5. **State Frequency Distributions of Refusing and Substitute Schools (For the states where one or more substitutions were required)**			
Measure	Greater in Substitute Schools	No Difference	Greater in Refusing Schools
Mean Eighth Grade Enrollment	8	0	6
Mean Percentage of black Students	6	3	5
Mean Percentage of Hispanic Students	5	3	6

Finally, it was possible to compare performance scores between schools in the original sample and substitute schools. The results, for 3,798 sessions in original schools and 117 sessions in substitute schools, are presented in Table 6. The mean scores in substitute schools were slightly higher, which merely reflects the fact that the substitute schools matched the refusing schools in being smaller and having fewer black students than cooperating schools, on the average. (The correlation across participating schools between eighth grade enrollment and mean TSA performance scores was -.12, $p < .0001$.)

Table 6. Means and Variances of Performance Scores for Originally Selected and Substitute Schools

Standardized Logit Scores	Not Adjusted for Ethnic and Sex Effects	Adjusted for Ethnic and Sex Effects
Mean for Original Schools	-.003	-.001
Mean for Substitute Schools	.106	.022
Within-Session Variance for Original Schools	.801	.766
Within-Session Variance for Substitute Schools	.824	.790
Between-Session Variance for Original Schools	.141	.077
Between-Session Variance for Substitute Schools	.120	.100
Between-state Variance for Original Schools	.091	.033
Between-state Variance for Substitute Schools	.047	.007
t-value for mean difference	-1.19	-0.40

Student participation rates. In each participating school, approximately 30 students were sampled for the TSA. According to Westat's preliminary report in April 1990, a total of 118,866 students were sampled, of whom 100,858 students assessed, including 99,195 in initial samples and 1,663 in make-up sessions. (If more than four students were missing from the session, the school was required to schedule a make-up session). Westat reported that 5,708 students were identified as withdrawn (i.e., not in eighth grade at the selected school), and 6,423 others did not show up for the testing. Excluded for prescribed reasons were 5,877 students. The completion rate within participating schools, for those who should have been assessed, was 94 percent.

Rates of student withdrawal. According to a data file provided by National Computing Systems (NCS), the subcontractor with responsibility for scoring the NAEP assessment forms, number of withdrawals was 5,710, extremely close to the number reported earlier by Westat. This number represents 5 percent of the total planned student sample, ranging from 2 percent in one state to 8 percent in another. Differences in withdrawal rates between states might be expected to have only a minimal effect on aggregate outcomes. However, there was a significant negative correlation, -.27, between the number of withdrawn students assigned to a session and the average performance score of those present in that session. That is, students tended to be withdrawn from sessions in which assessed students performed more poorly than those in other sessions. (The correlation was -.17 after score differences associated with race and sex were removed.)

Rates of student exclusion. Westat reported that approximately 5 percent of the sampled students were excluded from the TSA (typically about 5 percent of the students are excluded from the national NAEP). Four and one tenth percent were excluded as IEP (out of 8.3 percent identified as IEP). Seven tenths of 1 percent were excluded as LEP (out of 1.4 percent identified as LEP).

A student identified as LEP was excluded from the assessment if the student was a native speaker of a language other than English, had been enrolled in an English-speaking school (not including a bilingual education program) for less than two years, and was judged to be incapable of taking part in the assessment.

A student identified as having an IEP or equivalent classification was excluded from the assessment if the student was mainstreamed less than 50 percent of the time in academic subjects and was judged incapable of participating meaningfully in the assessment, or the IEP team or equivalent group had determined that the student was incapable of participating meaningfully in the assessment.

When there was doubt, LAs were told to include the student. Students were initially identified as IEP or LEP from school records; then the decision on exclusion was made according to the NAEP criteria. Thus, the rates for identification and exclusion could be different: some IEP and LEP students were tested.

Rates of student absenteeism. According to a data file provided by NCS, the number of absentees was 8,122, of whom 1,667 were tested in make-up sessions. The number of absentees represents 7 percent of the total sample, ranging from 4 percent in one state to 18 percent in another. The state with the 18 percent absenteeism rate clearly differs from the remainder of the sample. For example, of a total of 46 sessions with 10 or more absences, 18 were in this one state. As in the case of withdrawals, there was a significant negative correlation, -.27, between the number of absent students assigned to a session and the average performance score of those present in that session. That is, students tended to be absent from sessions in which assessed students performed more poorly than those in other sessions. (The correlation was -.18 after score differences associated with race and sex were removed.)

Questionnaire response rates. Finally, the response rates on auxiliary instruments are to be considered. These included a questionnaire on each schools' characteristics and policies, a questionnaire on each math class taken by an examinee completed by the math teacher, and a questionnaire for each formally excluded student. According to Westat's preliminary report, 99 percent of these instruments were completed.

Validity Issues Related to Materials Transmission

Irregularities in the reproduction and transmission of materials can affect the uniformity of the administration in various ways. If materials fail to reach a site in time for the administration or insufficient materials are sent, students are inadvertently lost from the sample (unless make-up sessions are held). The same is true if completed answer sheets are lost before they are returned for scoring. Also improper spiraling of test booklets--either in the shipment of materials or in their distribution on site--is a threat to the validity of the matrix sampling design. If materials are defective (improperly printed booklets, malfunctioning calculators), then the test situation is degraded for the students involved, and they have less opportunity to demonstrate their competence. Finally, if security is breached and some students have the opportunity to view the test materials ahead of time (or teachers view the materials and provide the students with relevant coaching), then these students are unfairly advantaged in the testing situation.

The first step was to examine responses to the QCM's pre-assessment telephone call. These questions were asked whether or not the session was to be monitored, and they provide very rough indications of the extent to which preparations were smooth, both in terms of receipt of materials from the NAEP contractor and in terms of the LA's completion of certain pre-assessment tasks. In Tables 7, 9, 11, 13, and 15, the frequencies of responses to five pre-assessment questions are presented. Not surprisingly, the items that were primarily the responsibility of the contractor showed very few exceptions (roughly 3 percent), whereas the items that were primarily the responsibility of the LA were more frequently not completed. A major purpose of the QCM call was to ensure that the pre-assessment tasks were taken care of; therefore, it is likely that those who had not completed tasks prior to the pre-assessment call completed them soon after.

The primary interest here is in whether these items might be indicators of general freedom from exceptions and smooth and error-free administration of the assessment sessions. Would scores be lower, or more variable, in sessions where there were indications of pre-assessment unpreparedness? The results in Tables 8, 10, 12, 14, and 16 suggest that assessment performance scores were not affected by pre-assessment variations. If anything, scores were slightly higher in sessions conducted by LAs who admitted to neither having carried out supplementary sampling nor having distributed teacher questionnaires prior to the QCM's call. The variations in these five tables, which are not significant, provide a background for assessing the size of real differences that occurred in scores as a function of variations in administration, discussed in the next section.

Table 7. Frequencies of Schools Who Did and Did Not Receive Administration Schedules prior to the Check-up Call

Status, according to QCM	Number of Sessions	Percentage of Sessions
Received On-Time	3495	97.3%
Received Late	71	2.0%
No Response	27	0.7%

Table 8. Means and Variances of Performance Scores for Sessions Receiving Administration Schedules on Time or Late

Standardized Logit Scores	Not Adjusted for Ethnic and Sex Effects	Adjusted for Ethnic and Sex Effects
Mean for Late Receipt	-.110	-.050
Mean for On-Time Receipt	.002	.001
Within-Session Variance for Late Receipt	.760	.717
Within-Session Variance for On-Time Receipt	.802	.767
Between-Session Variance for Late Receipt	.125	.044
Between-Session Variance for On-Time Receipt	.141	.078
Between-State Variance for Late Receipt	.117	.051
Between-State Variance for On-Time Receipt	.089	.032
t-value for mean difference	-1.37	-0.89

Table 9. Frequencies of Schools Who Did and Did Not Complete Supplementary Student Sampling prior to the Check-up Call		
Status, according to QCM	Number of Sessions	Percentage of Sessions
Completed	2973	84.4%
Not Completed	494	14.0%
No Response	55	1.6%

Table 10. Means and Variances of Performance Scores for Sessions in which Administrators Completed Supplementary Sampling on Time or Late		
Standardized Logit Scores	Not Adjusted for Ethnic and Sex Effects	Adjusted for Ethnic and Sex Effects
Mean for Late Sampling	.052	.027
Mean for On-Time Sampling	-.007	-.004
Within-Session Variance for Late Sampling	.768	.733
Within-Session Variance for On-Time Sampling	.806	.771
Between-Session Variance for Late Sampling	.128	.061
Between-Session Variance for On-Time Sampling	.142	.080
Between-State Variance for Late Sampling	.069	.021
Between-State Variance for On-Time Sampling	.093	.033
t-value for mean difference	1.73	1.34

Table 11. Frequencies of Schools Who Did and Did Not Receive Assessment Materials prior to the Check-up Call

Status, according to QCM	Number of Sessions	Percentage of Sessions
Received	3420	95.2%
Not Received	111	3.1%
Missing	61	1.7%

Table 12. Means and Variances of Performance Scores for Sessions in which Assessment Materials Received on Time or Late

Standardized Logit Scores	Not Adjusted for Ethnic and Sex Effects	Adjusted for Ethnic and Sex Effects
Mean for Late Receipt	-.060	-.020
Mean for On-Time Receipt	.001	.001
Within-Session Variance for Late Receipt	.768	.709
Within-Session Variance for On-Time Receipt	.802	.768
Between-Session Variance for Late Receipt	.148	.089
Between-Session Variance for On-Time Receipt	.141	.078
Between-State Variance for Late Receipt	.134	.038
Between-State Variance for On-Time Receipt	.088	.032
t-value for mean difference	-0.78	-0.40

Table 13. Frequencies of Schools Who Did and Did Not Receive Sufficient Quantities of Assessment Materials prior to the Check-up Call

Status, according to QCM	Number of Sessions	Percentage of Sessions
Sufficient	2693	77.4%
Insufficient	165	4.7%
Not Checked	553	15.9%
Missing	70	2.0%

Table 14. Means and Variances of Performance Scores for Sessions in which Assessment Materials Received Were Incomplete

Standardized Logit Scores	Not Adjusted for Ethnic and Sex Effects	Adjusted for Ethnic and Sex Effects
Mean when Incomplete	-.019	.061
Mean when Complete	.001	-.003
Within-Session Variance when Incomplete	.853	.799
Within-Session Variance when Complete	.799	.765
Between-Session Variance when Incomplete	.173	.082
Between-Session Variance when Complete	.139	.077
Between-State Variance when Incomplete	.060	.019
Between-State Variance when Complete	.091	.033
t-value for mean difference	-0.42	1.91

Table 15. Frequencies of Schools Who Did and Did Not Distribute Teacher Questionnaires prior to the Check-up Call

Status, according to QCM	Number of Sessions	Percentage of Sessions
Distributed	2225	64.9%
Not Distributed	1138	33.2%
Missing	67	1.9%

Table 16. Means and Variances of Performance Scores for Sessions in which Teacher Questionnaires were Distributed On-Time or Late

Standardized Logit Scores	Not Adjusted for Ethnic and Sex Effects	Adjusted for Ethnic and Sex Effects
Mean when Late	.032	.016
Mean when On-Time	-.013	-.006
Within-Session Variance when Late	.801	.763
Within-Session Variance when On-Time	.802	.768
Between-Session Variance when Late	.136	.067
Between-Session Variance when On-Time	.141	.082
Between-State Variance when Late	.057	.023
Between-State Variance when On-Time	.103	.036
t-value for mean difference	1.56	1.22

Validity Issues Related to Session Administration

Irregularities in session administration affect validity primarily by creating especially favorable or unfavorable testing conditions for particular students. For example, inaccuracies in the timing of subtests could result in some students having more or less time than intended to complete certain testing tasks; inadequate supervision by testing monitors could provide unfair opportunities to copy from the work of others or to spend additional time working on test sections. Malfunctioning calculators, inadequate instruction in their use, or failure to distribute them to all of the proper students could invalidate test questions that were intended to be answered with the aid of a calculator. Furthermore, improper distribution of the calculators may have given an unfair advantage to students who had access to calculators in the non-calculator blocks of the assessment. Deviations by the LAs in reading instructions might create confusion or misunderstandings among some students so that they could not properly execute the testing tasks. On the other hand, unsanctioned elaborations on instructions, or pep talks, might provide an unfair advantage. Finally, testing late in the day when fatigue might be a factor, substandard room conditions (light, heat, noise), or general lack of discipline among the students could unfavorably influence the performance of certain groups of students.

For most of these potential problems, information on the *frequency of occurrence* is available on the forms filled out by the QCMs, and generally these data are quite complete, with fewer than 1 percent missing responses on most items. However, since the QCMs were instructed to intervene to correct observed problems wherever possible, there is only limited opportunity to assess the probable magnitude of the *impact on student scores* in the absence of such intervention. A related but somewhat lesser problem is that, with the current file structure, score data for students who attended makeup sessions cannot be disaggregated from the average session scores, even though they were not tested under the same conditions.

The QCM forms provide information on the following aspects of LA performance that are pertinent to the adequacy of the session administration:

- timing of booklet sections,
- instruction on use of calculators,
- distribution and collection of calculators,

- reading of test instructions (session script),
- handling of student questions,
- monitoring of student work, and
- overall evaluation of the LA's performance.

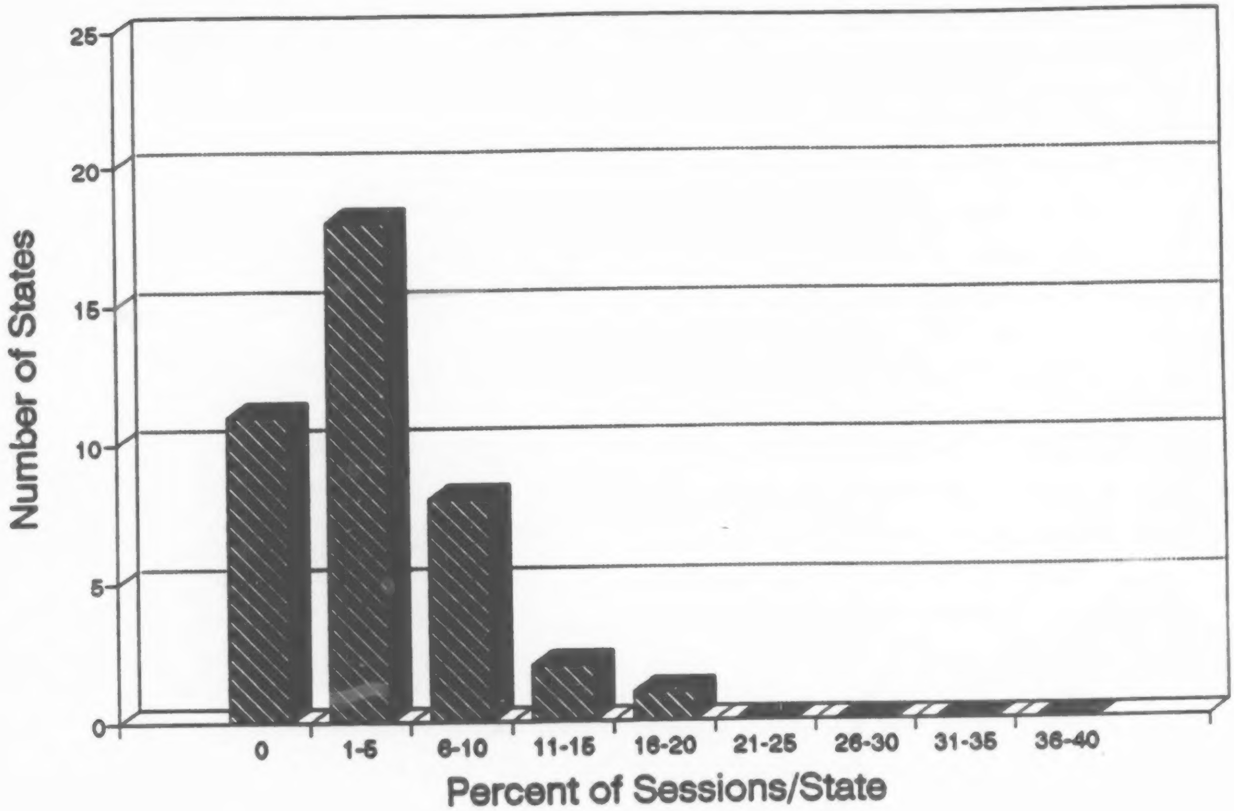
In addition, information is available on a few conditions of the administration that may or may not be under the control of the LA, but that also have the potential to affect student performance. These factors include:

- the time of day of the assessment,
- the adequacy of the room used for the assessment, and
- the attitude and behavior of the students participating in the assessment.

Timing of booklet sections. Student booklets contained five separately timed sections: two background sections followed by three cognitive sections. The majority of the timing errors occurred in the first section, suggesting that some of the LAs needed practice to master this aspect of their assignment. The placement of the background sections (where timing is not critical) ahead of the cognitive sections was therefore a good strategy for minimizing the potential impact of timing errors made by novice test administrators.

Timing errors were reported in 3 percent of the monitored assessment sessions, and Figure 1 shows the distribution of these sessions across states. Eleven states had no sessions with timing errors, 18 states had timing errors in 1 percent to 5 percent of sessions, and only 1 state had timing errors in more than 15 percent of sessions. No more than two timing errors were reported in any one session, while the QCMs reported making interventions in connection with 86 percent of the reported errors.

Figure 1. One or More Errors Timing Sections



As can be seen in Table 17, mean performance scores adjusted for ethnicity and sex tend to decline with increasing numbers of timing errors; the mean adjusted performance scores for sessions with one or more timing errors was -0.080 compared to 0.013 for the remainder (t = -1.92).

Table 17. Means and Variances of Adjusted Performance Scores by Number of Errors in Timing (N = 1,918 Sessions)

No. of Errors	0	1	2
No. of Sessions	1852	60	6
Mean	.013	-.071	-.182
Within-Session Variance	.768	.699	.614
Between-Session Variance	.076	.095	-[1]
Between-State Variance	.031	.007	-[1]

[1] Insufficient data to estimate

The exact causal relationship between irregularities in section timing and student performance is not entirely clear, however. If cognitive sections were allowed to run longer than intended, one might expect an *increase* in student performance, while, on the other hand, if sections were cut short or otherwise disrupted by timing errors, one might expect a *decrease*. Currently, coding of QCM comments concerning the specific nature of the timing irregularities is not available. However, as Table 18 shows, it is the timing of the *background* sections, rather than of the cognitive sections, that exhibits the strongest relationship to student performance.

Table 18. Means and Variances of Adjusted Performance Scores for Specific Timing Errors (1,918 Sessions)

Error in Timing for	Section 1 (Backgrd)	Section 2 (Backgrd)	Section 3 (Cognitive)	Section 4 (Cognitive)	Section 5 (Cognitive)
No. Sessions with Error	42	7	9	7	7
Mean with Error	-.134	-.215	.090	-.068	.056
Mean w/out Error	.013	.011	.010	.011	.010
Within-Session Variance with Error	.664	.578	.852	.661	.731
Within-Session Variance w/out Error	.768	.766	.765	.766	.766
Between-Session Variance with Error	.045	-[1]	-[1]	-[1]	-[1]
Between-Session Variance w/out Error	.076	.076	.075	.076	.075
Between-State Variance with Error	.040	-[1]	-[1]	-[1]	-[1]
Between-State Variance w/out Error	.031	.031	.031	.032	.032
t-value for mean difference	-2.31*	-[1]	-[1]	-[1]	-[1]

[1] Insufficient data to estimate

* Significant at $p \leq .05$

Instruction on use of calculators. Each session began with instruction on use of the calculators provided. The QCM evaluated the LA's performance on four aspects of this task: correctly covering the verbal directions (as given in the script), using the poster consistently during the presentation to point to the various calculator keys, allowing sufficient time for all students to complete the practice exercises, and answering questions appropriately during the training.

Table 19 shows the relationship of student performance to the number of irregularities reported in calculator training. As can be seen, two or three irregularities are indicative of lower student performance, but a single irregularity is not. However, as with timing errors, the relationship fails to reach statistical significance; the mean adjusted performance score for sessions with two or more irregularities was -0.123 compared to 0.012 for the remainder (t = -1.80).

Table 19. Means and Variances of Adjusted Performance Scores by Number of Errors in Calculator Training (N = 1,918 Sessions)

No of Errors	0	1	2	3	4
N	1709	187	24	2	1
Mean	.009	.042	-.140	-.401	.616
Within-Session Variance	.768	.746	.755	.702	.593
Between-Session Variance	.075	.071	-[1]	-[1]	-[1]
Between-State Variance	.032	.025	-[1]	-[1]	-[1]

[1] Insufficient data to estimate

Two or more irregularities occurred in only 1 percent of the assessment sessions; Figure 2 shows how these sessions were distributed across the states. Twenty-three states had no sessions with two or more irregularities, 14 states had such irregularities in 1 percent to 5 percent of sessions, and 3 states had them in 6 percent to 10 percent of sessions.

As shown in Table 20, the most common irregularity, which affected 7.4 percent of sessions, was failure to use the poster consistently in delivering calculator training. The other error rates include inadequacies in delivering the verbal instructions (3.3 percent of sessions); failure to complete student practice (1.3 percent of sessions); and finally, inadequacies in answering student questions (0.7 percent of sessions). QCM interventions were reported for 83 percent of the irregularities in verbal instructions, 68 percent of the failures to complete student practice, and 85 percent of the problems with student questions. There was no indication that QCMs were instructed to intervene to ensure proper use of the poster.

Irregularities with regard to poster use or student practice were associated with somewhat *higher* mean performance scores, although none of the differences are statistically significant. One might speculate that LAs moved more quickly through the training in schools where students were already well versed in calculator use, and that such schools tended to exhibit higher levels of math achievement. *Lower* scores were associated with irregularities in delivering the verbal instructions and answering student questions, despite the high rate of QCM interventions (which would tend to offset the effects of such errors).

Figure 2. Two or More Errors in Calculator Training

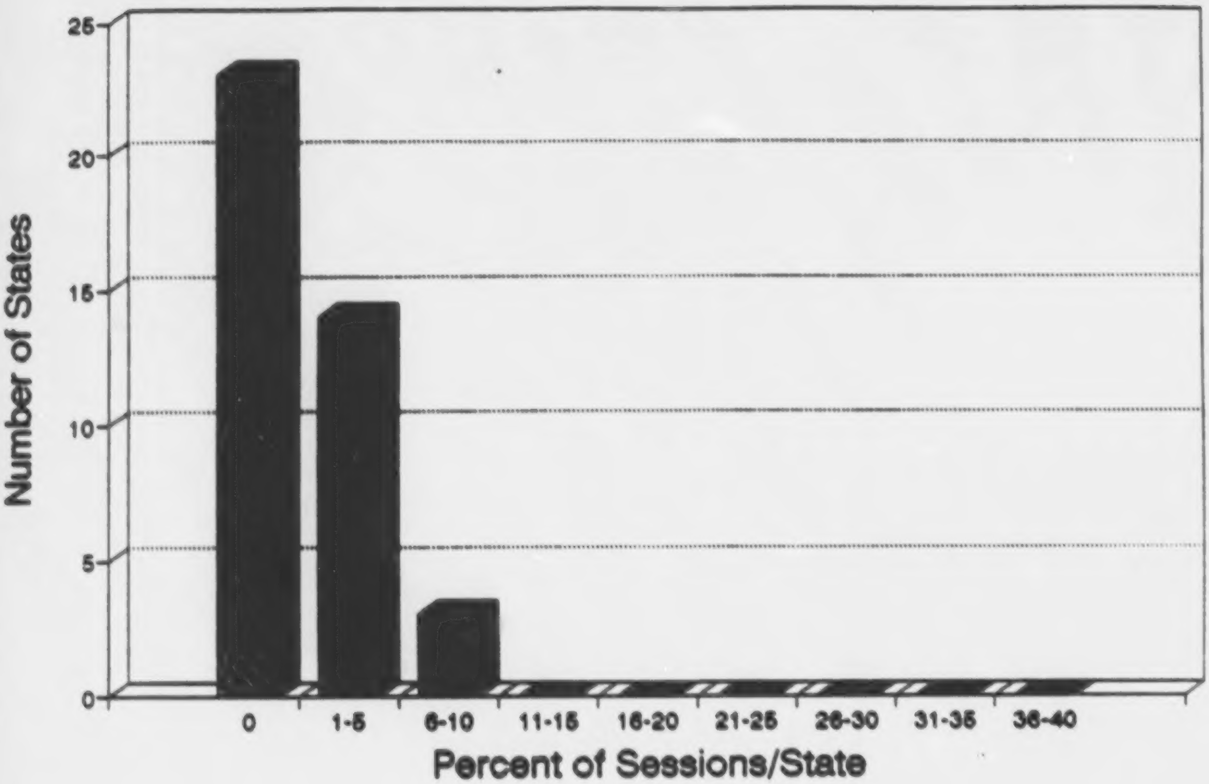


Table 20. Means and Variances of Adjusted Performance Scores for Specific Calculator Training Errors (N = 1,918 Sessions)

Error in	Using Poster	Delivering Instructions	Completing Student Practice	Answering Questions
No. Sessions with Error	141	64	17	13
Mean with Error	.034	-.079	.121	-.112
Mean w/out Error	.008	.013	.009	.011
Within-Session Variance with Error	.766	.728	.707	.658
Within-Session Variance w/out Error	.765	.767	.766	.766
Between-Session Variance with Error	.086	.063	.079	-[1]
Between-Session Variance w/out Error	.075	.076	.076	.076
Between-State Variance with Error	.029	.038	.023	-[1]
Between-State Variance w/out Error	.032	.032	.031	.031
t-value for mean difference	0.74	-1.77	1.50	-[1]

[1] Insufficient data to estimate

Distribution and collection of calculators. In order to test math skills using both calculators and hand calculations, calculators had to be distributed and picked up between booklet sections. Moreover, in order to maximize the use of an expensive resource, the sections were staggered so that different students would be using the calculators at different points in time. While this staggered distribution had definite cost advantages, it did increase the complexity of the LA's assignment. In fact, errors in calculator distribution and/or collection were the second most common irregularity, occurring in 11 percent of the monitored sessions. As shown in Figure 3, the median state had errors in calculator distribution in 6 percent to 10 percent of sessions, and two states had such errors in 38 percent to 40 percent of sessions.

Calculators were distributed and collected four times during each session: before and after the training and before and after each of the three cognitive sections. However, the QCMs noted adequacy of performance for only seven of these instances (skipping distribution before training), so the maximum number of irregularities that could be recorded for a given session was seven. Table 21 shows that four or more errors were recorded in a few sessions, but in 98 percent of the sessions there were either no errors or one error. Moreover, there is no relationship between the number of errors and the average performance score for the session. This may be due to the fact that, out of the 283 separate instances of distribution errors, QCM intervention was recorded in all but 11 instances (96 percent).

The highest proportion of errors (47 percent) occurred in the distribution of calculators for the first of the cognitive sections. Since the distribution for calculator training was simply to every other student, this was the first time that the LAs had to coordinate the distribution with the code on the student booklets that indicated whether a student was supposed to have a calculator for the upcoming section.

Figure 3. **One or More Errors in Calculation Distribution**

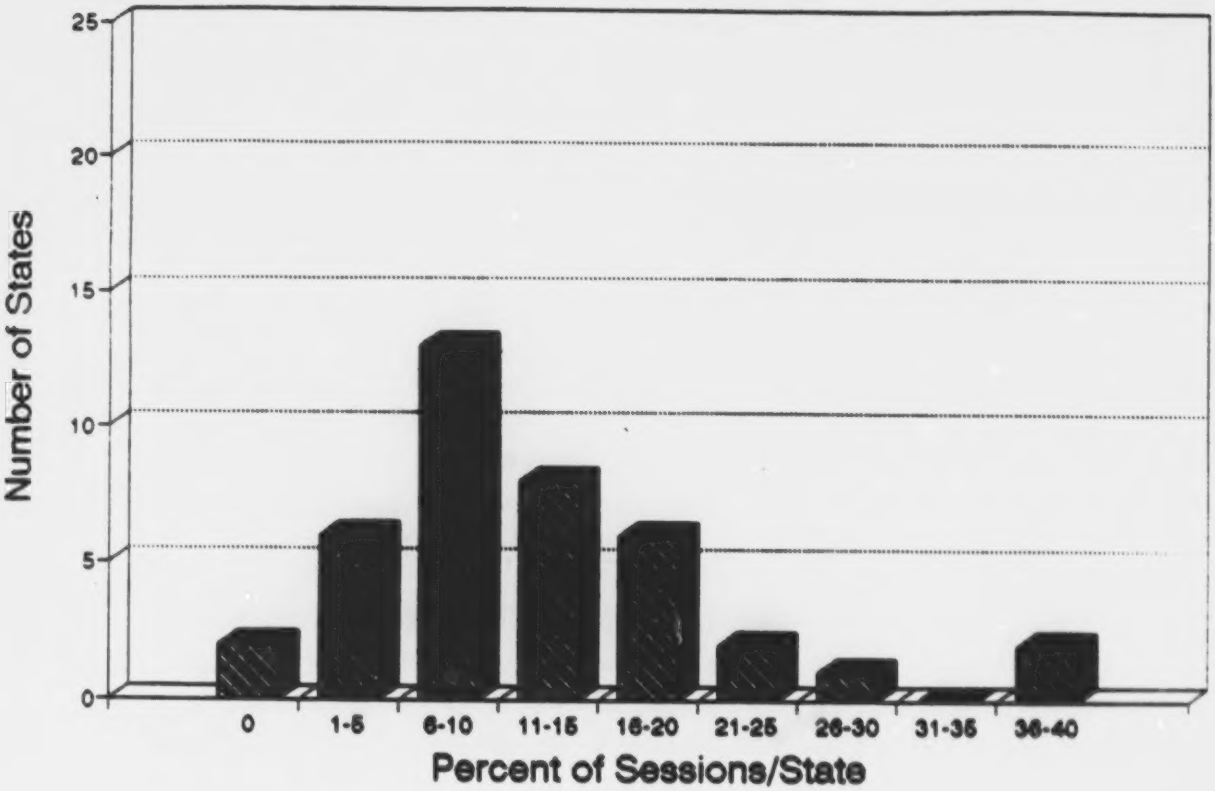


Table 21. **Means and Variances of Adjusted Performance Scores by Number of Errors in Distribution and Collection of Calculators (N = 1,918 Sessions)**

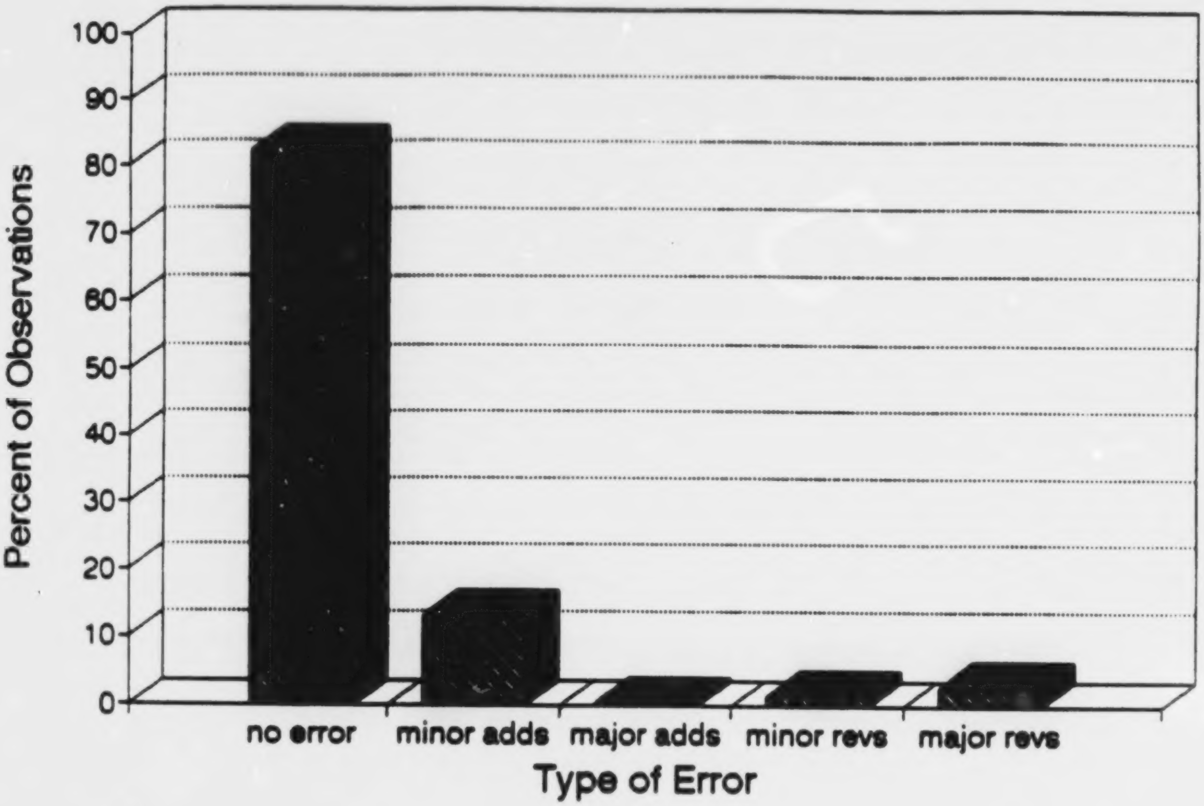
No of Errors	0	1	2	3	4	5	6	7
N	1698	182	28	5	1	1	2	1
Mean	.014	-.030	-.001	.263	.163	.254	.099	-.291
Within-Session Variance	.768	.737	.785	.712	1.36	1.28	.634	-[1]
Between-Session Variance	.072	.112	.002	-[1]	-[1]	-[1]	-[1]	-[1]
Between-State Variance	.031	.039	.101	-[1]	-[1]	-[1]	-[1]	-[1]

[1] Insufficient data to estimate

Reading of test instructions. LAs were provided with a script that covered all instructions required to guide students through the assessment, and they were asked to confine themselves to the script in order to maintain uniformity of administration. QCMs scored LA performance on five script segments: (1) introduction and distribution of materials, (2) when to use the calculators, (3) how to code the booklet covers, (4) how to use the test booklets and indicate answer choices, and (5) beginning the assessment and responding to the first question (racial-ethnic identification).

In all, 1,690 coded irregularities occurred in script reading, comprising 17.7 percent of the rated script elements. As can be seen in Figure 4, minor additions were the most frequent type of irregularity (13.2 percent); followed by major revisions or omissions (2.8 percent), minor revisions or omissions (1.3 percent), and major additions (0.4 percent). QCM intervention was indicated in 223 of the 264 instances of major revisions or omissions (84 percent).

Figure 4. Types of Errors Observed Reading Script



(Five observations per session)

Although the majority of script-reading irregularities were only minor additions and QCMs intervened in the majority of instances of major revisions or omissions, there was still a statistically significant relationship between the number of script reading errors in a session and mean student performance. As shown in Table 22, the mean adjusted performance score was below the overall mean for sessions with three, four, or five irregularities. (The mean adjusted performance score for sessions with three or more irregularities was -0.084 compared to 0.024 for the remainder; $t = -3.42$ $p \leq .001$.) This might suggest that the net effect of such irregularities was to depress student performance by creating confusion. A competing hypothesis, however, is that LAs deviated from the script to clarify the assessment procedures for low performing students.

Table 22. Means and Variance of Adjusted Performance Scores by Number of Errors in Reading Script (1,918 Sessions)

No of Errors	0	1	2	3	4	5
N	1097	399	178	107	79	58
Mean	.025	.008	.055	-.034	-.120	-.127
Within-Session Variance	.770	.766	.807	.736	.677	.731
Between-Session Variance	.069	.099	.089	.063	.073	.054
Between-State Variance	.028	.031	.013	.033	.052	.071

Table 23 shows that lower student performance is associated with irregularities in *any portion of the script*. The relationship is statistically significant for the script sections on when to use the calculators, how to use the test booklets, and beginning the assessment.

Table 23. Means and Variances of Adjusted Performance Scores for Specific Script Reading Errors (N = 1,918 Sessions)

Error in Script for	Intro/ Distrib of Materials	When to Use Calculators	Coding Booklet Covers	Using the Test Booklet	Beginning Assess/ Backgrnd Section
No. Sessions with Error	411	304	297	367	303
Mean with Error	-.022	-.038	-.026	-.060	-.041
Mean w/out Error	.019	.019	.017	.027	.020
Within-Session Variance with Error	.769	.749	.751	.727	.733
Within-Session Variance w/out Error	.764	.769	.768	.775	.772
Between-Session Variance with Error	.073	.081	.075	.074	.078
Between-Session Variance w/out Error	.075	.075	.076	.075	.076
Between-State Variance with Error	.052	.045	.046	.035	.039
Between-State Variance w/out Error	.027	.029	.028	.031	.029
t-value for mean difference	-1.66	-2.20*	-1.54	-3.21**	-2.13*

* Significant at $p \leq .05$
** Significant at $p \leq .01$

Three or more script irregularities were observed in 13 percent of the monitored sessions; Figure 5 shows how these sessions were distributed across the states. Only one state had no sessions with three or more irregularities, and the median state had such irregularities in 11 percent of sessions.

Handling student questions. The QCMs reported that LAs handled student questions inappropriately in only 1 percent of sessions, distributed across the states as shown in Figure 6. However, this irregularity *was* significantly associated with lower student performance (Table 24). The mean adjusted performance score for sessions in which student questions were handled inappropriately was -.184, compared to .012 for all others; $t = 2.11$ $p \leq .05$.

The available coding does not provide substantive explanations for all irregularities involving student questions (N=17). On the basis of available data, it appears that problems included both behaviors that might be expected to *raise* scores (e.g., giving unsanctioned instructions to students), and behaviors that might *lower* scores (e.g., ignoring raised hands and speaking too softly to be heard). This explanation is consistent with the unusually high between-session variance observed for these sessions.

Figure 5. Three or More Errors Reading Script

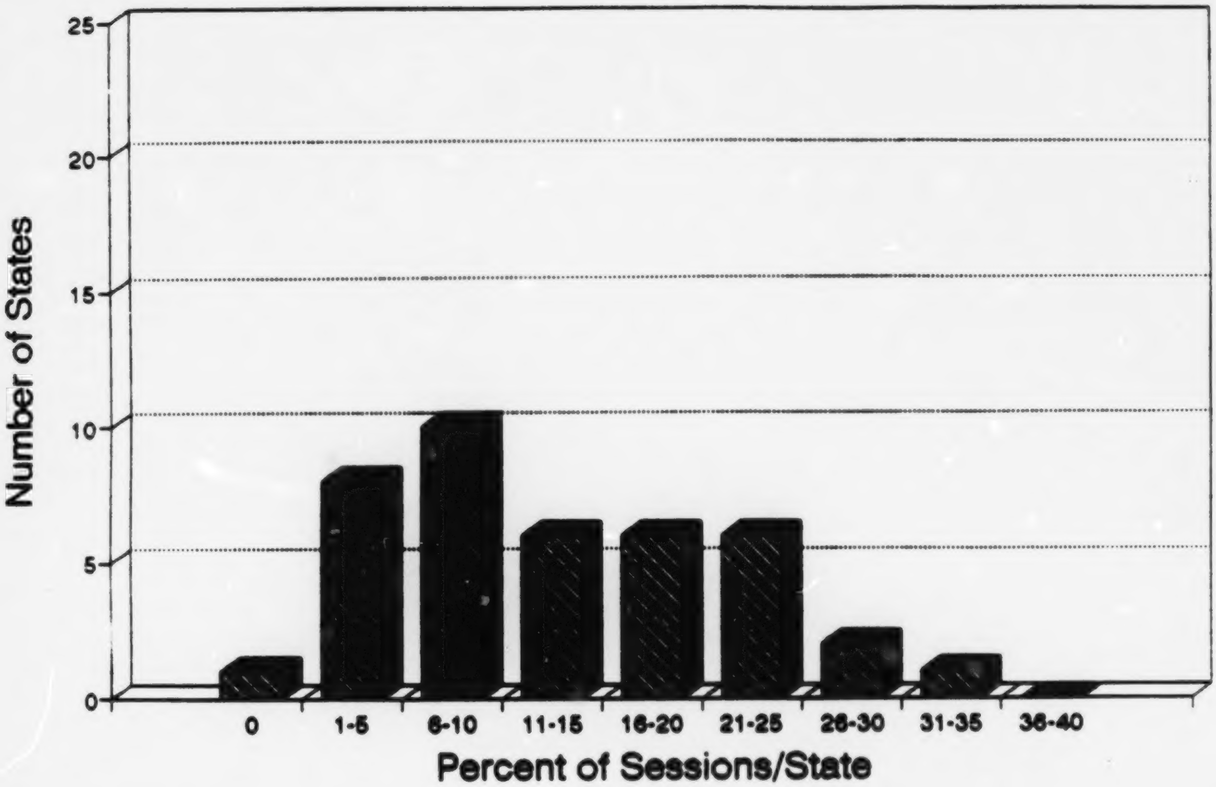


Figure 6. Improper Handling of Student Questions

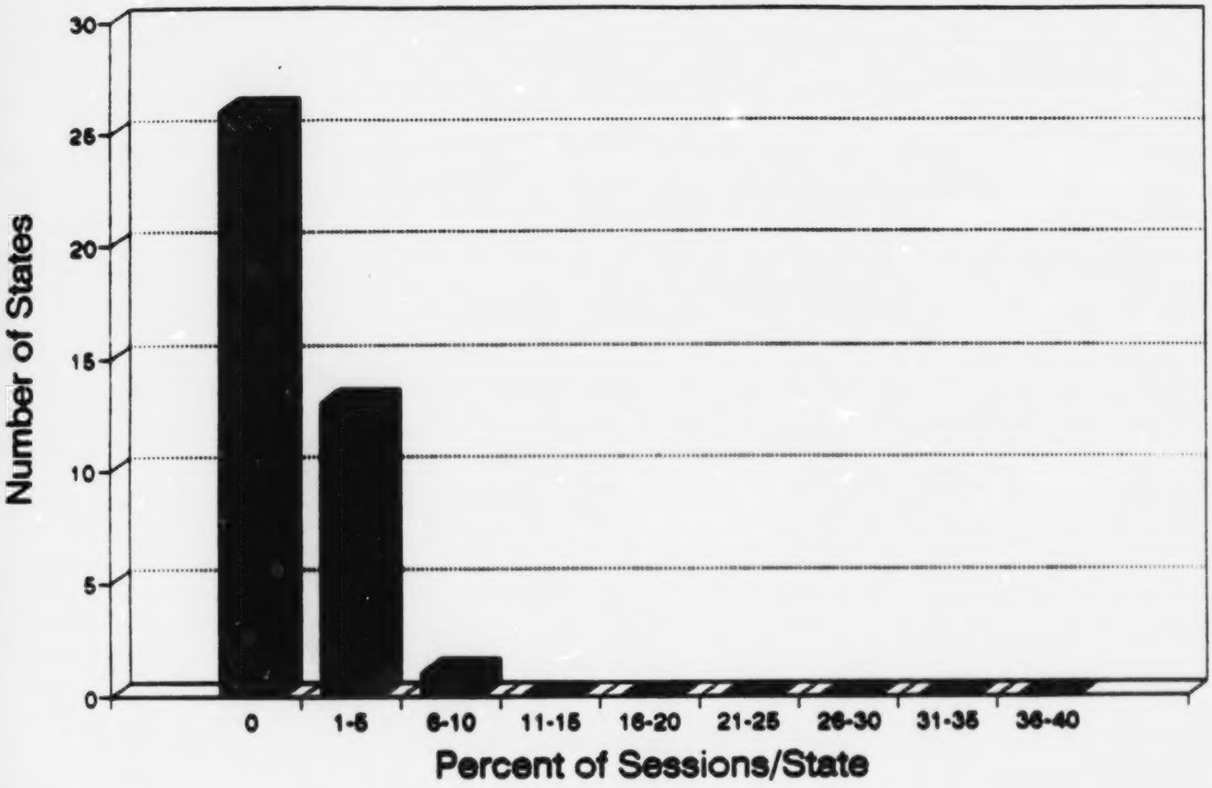


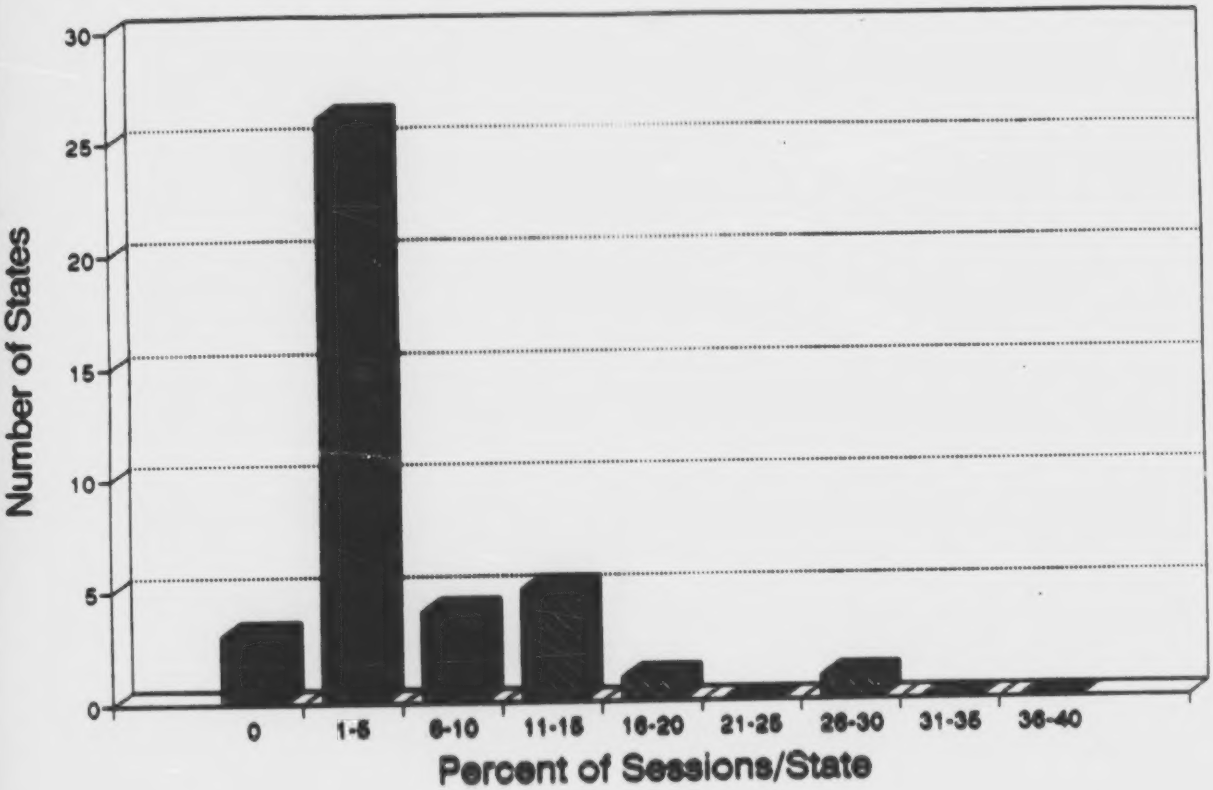
Table 24.	Means and Variances of Adjusted Performance Scores Inappropriate Handling of Student Questions (N = 1,918 Sessions)				
	No. of Sessions	Mean	Within-Session Variance	Between-Session Variance	Between-state Variance
With Problem	17	-.184*	.791	.127	.059
Without Problem	1901	.012	.765	.075	.031

* Significant at $p \leq .05$

Monitoring student work. Copying is not likely to be ajor threat to validity in an assessment with a spiralled design, such as the T because most students will not be seated next to other students working in the same booklet. However, students could still obtain an unfair advantage by soliciting help from their neighbors, by working on sections before or after the time allotted, or by using their calculators to go back over questions in sections intended to test proficiency in hand calculation. Conversely, students could affect their scores in an unknown manner if they became confused and worked the sections out of order, or disregarded the intended timing. The LAs were expected to monitor students' work to avert such problems, and the QCMs were instructed to score one objective marker of the LA's performance in this area: whether or not the LA walked around the room to make sure that students were working in the correct sections.

LAs failed to monitor by walking around in 5 percent of the monitored sessions, not counting nine instances in which the QCM noted that it was not physically possible for the LA to walk around. As shown in Figure 7, this problem occurred in 1 percent to 5 percent of sessions for most states, although one state had this problem in 29 percent of sessions.

Figure 7. **Failure to Monitor by Walking Around**



There is no indication that, on the average, students benefitted from this lack of vigilance--mean adjusted performance scores were somewhat lower in the sessions where the LA failed to walk around, although the difference is not statistically significant (Table 25; t = -1.56).

Table 25. Means and Variances of Adjusted Performance Scores Failure to Monitor by Walking Around (N = 1,918 Sessions)

	No. of Sessions	Mean	Within-Session Variance	Between-Session Variance	Between-state Variance
With Problem	104	-.051	.733	.073	.039
Without Problem	1814	.014	.767	.076	.031

Overall rating of local administrator performance. The QCMs also provided an overall rating of the LA's performance, using a five-point scale that ranged from 1, "LA needed no assistance," to 5, "could not have been done without you." The QCM rating covered responsibilities beyond the administration of the assessment session proper, including maintaining the roster of sampled students; distributing and collecting school, teacher, and excluded student questionnaires; and notifying teachers and students of the assessment schedule. Nevertheless, there were significant correlations between the overall ratings given to the LAs and the various irregularities observed in the sessions; these are shown in Table 26.

Table 26. Correlations between Rating of Overall Local Administrator Performance and Various Irregularities in Session Administration

	Number of Script Irregularities	Number of Timing Irregularities	Number of Irregularities in Calc. Training	Number of Irregularities in Calc. Distrib.
Correlation Coefficient	0.337	0.166	0.352	0.285
Probability	.0001	.0001	.0001	.0001

There was also a significant relationship between the overall rating received by the LA and average student performance. As can be seen in Table 27, ratings of three ("mixed results") or worse are associated with below-average mean performance scores. The mean adjusted performance score for sessions with rating of three or worse is -0.045 compared to 0.018 for the remainder, $t = -2.11$ $p \leq .05$. On the other hand, there was *no significant difference* in student performance based on the LA's perceived attitude towards the assessment. (Comparison was made between the 206 sessions where LAs were rated "mixed/indifferent" or "negative" and the 1,712 sessions where they were rated "positive.")

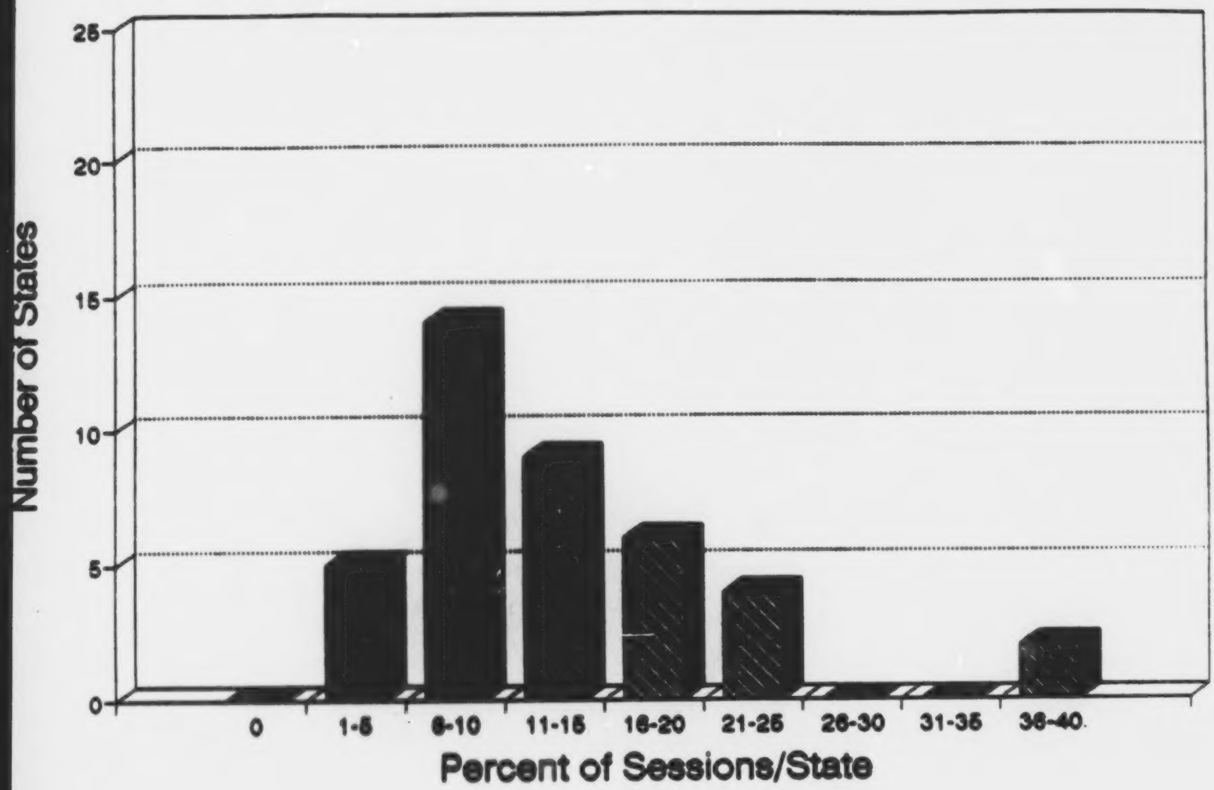
Overall LA ratings were *not* correlated with the SES of the school (as measured by average educational level of mothers or by percentage of students eligible for the federal lunch program). There was, however, a correlation of 0.229 ($p \leq .0001$) between the presence of student disruption and overall ratings of "mixed" or worse.

Ratings of three or worse were given in 13 percent of monitored sessions; the distribution of these sessions across states is shown in Figure 8. There was no state in which all LAs received ratings of two or better, and the median state had low-performing LAs in 11 percent of sessions. In two states, ratings of three or worse were given in 38 percent of sessions. Special attention should be given to the TSA procedures in those two states in an effort to determine what could be done to improve performance in the next administration.

Table 27. Means and Variances of Adjusted Performance Scores by Rating of Local Administrator's Overall Performance (N = 1,918 Sessions)

QCM Rating	Needed No Assistance (1)	(2)	Mixed Results (3)	(4)	QCM Help Essential (5)
N	1006	650	185	42	15
Mean	.023	.007	-.017	-.152	-.090
Within-Session Variance	.780	.750	.733	.766	.791
Between-Session Variance	.073	.079	.077	.091	.145
Between-State Variance	.032	.029	.035	-.006	-.015

Figure 8. Local Administrator Performance Rated Mixed or Worse



Other conditions of administration. As noted above, information is available on a few conditions of administration that may not have been under the control of the LA, but that also could have affected student performance. These include the time of day of the administration, the adequacy of the room conditions, and the attitude and behavior of the students participating in the assessment.

Table 28 shows the relationship between time of day and average student performance. As can be seen, most sessions began between 8 a.m. and 1 p.m., and none began earlier than 7 a.m. or later than 2 p.m. No sessions began very late in the day, when fatigue might be expected to be a significant factor affecting performance. Although average scores were lower in the few sessions that began at 2 p.m. (mean = -0.205), there is no clear evidence of a pattern of relationship between time of day and student performance.

Table 28. Means and Variances of Adjusted Performance Scores by Time of Day of Assessment (N = 1,918 Sessions)

Hour Script Begun	7 a.m.	8 a.m.	9 a.m.	10 a.m.	11 a.m.	12 a.m.	1 p.m.	2 p.m.
N	11	624	796	111	21	126	188	8
Mean	-.036	-.010	.036	-.065	.034	.007	.033	-.205
Within-Session Variance	.869	.762	.763	.732	.840	.818	.783	.615
Between-Session Variance	-[1]	.079	.078	.100	-[1]	.062	.062	-[1]
Between-State Variance	-[1]	.043	.026	.020	-[1]	.036	.024	-[1]

[1] Insufficient data to estimate

The QCMs rated assessment rooms as inadequate in terms of size and/or freedom from distraction and noise in 4 percent of sessions. Figure 9 shows how these sessions were distributed across states. Nine states had no sessions held in inadequate rooms, 18 states had inadequate rooms for 1 percent to 5 percent of sessions, 10 states had inadequate rooms for 6 percent to 10 percent of sessions, and only 2 states had inadequate rooms for more than 10 percent of sessions.

Table 29 shows that somewhat lower student performance was observed in sessions held in rooms that were judged to be inadequate, although the relationship fell short of statistical significance (t = -1.85). There was no relationship between the SES level of the school (as judged by mother's education or percent eligible for federal lunch program) and the adequacy of the testing facility.

Table 29. Means and Variances of Adjusted Performance Scores Sessions Held in Inadequate Rooms (N = 1,918 Sessions)

	No. of Sessions	Mean	Within-Session Variance	Between-Session Variance	Between-state Variance
With Problem	86	-0.65	.751	.081	.006
Without Problem	1832	.014	.766	.076	.032

As might be expected, some of the strongest associations to average student performance were found on the items that measured student attitudes towards the assessment and behaviors during the assessment. Disruptive student behavior was noted in 7 percent of sessions, distributed across states as shown in Figure 10. The median state experienced student disruption in 6 percent of sessions, and no state experienced it in more than 17 percent of sessions.

Figure 9. Sessions Held in Inadequate Rooms

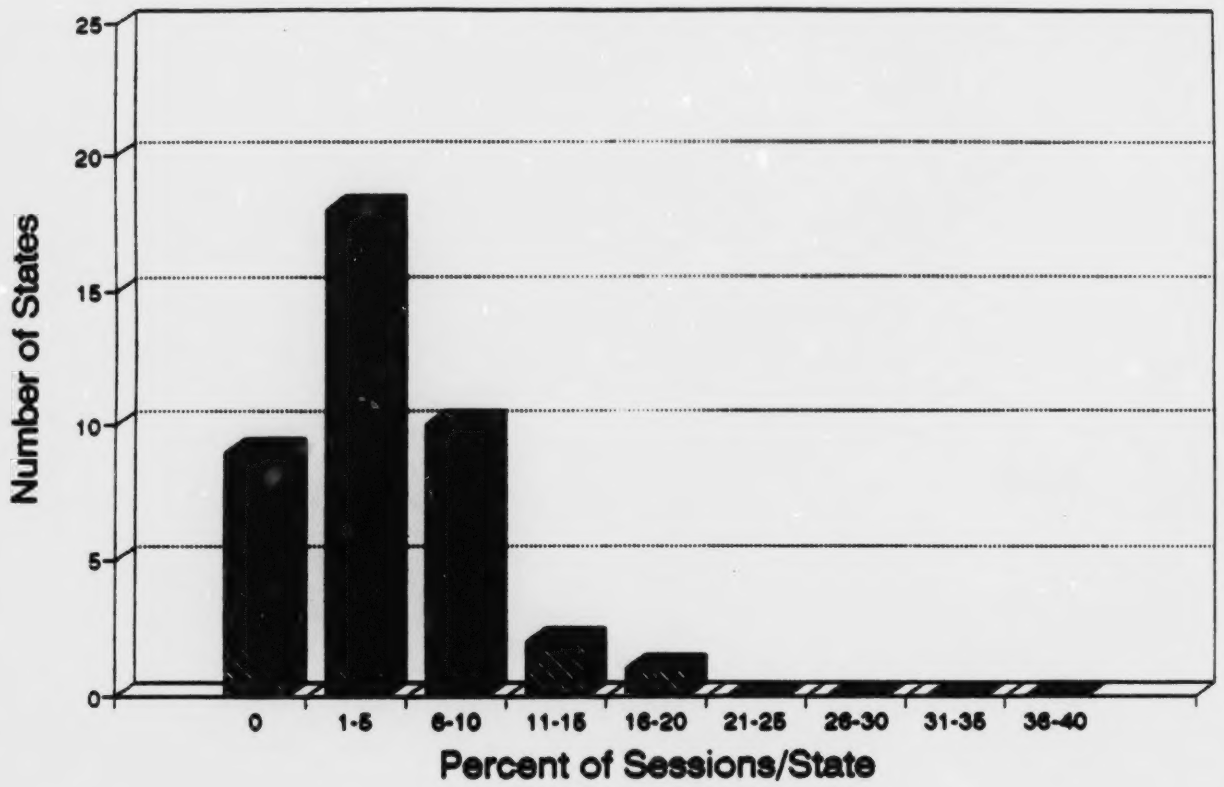
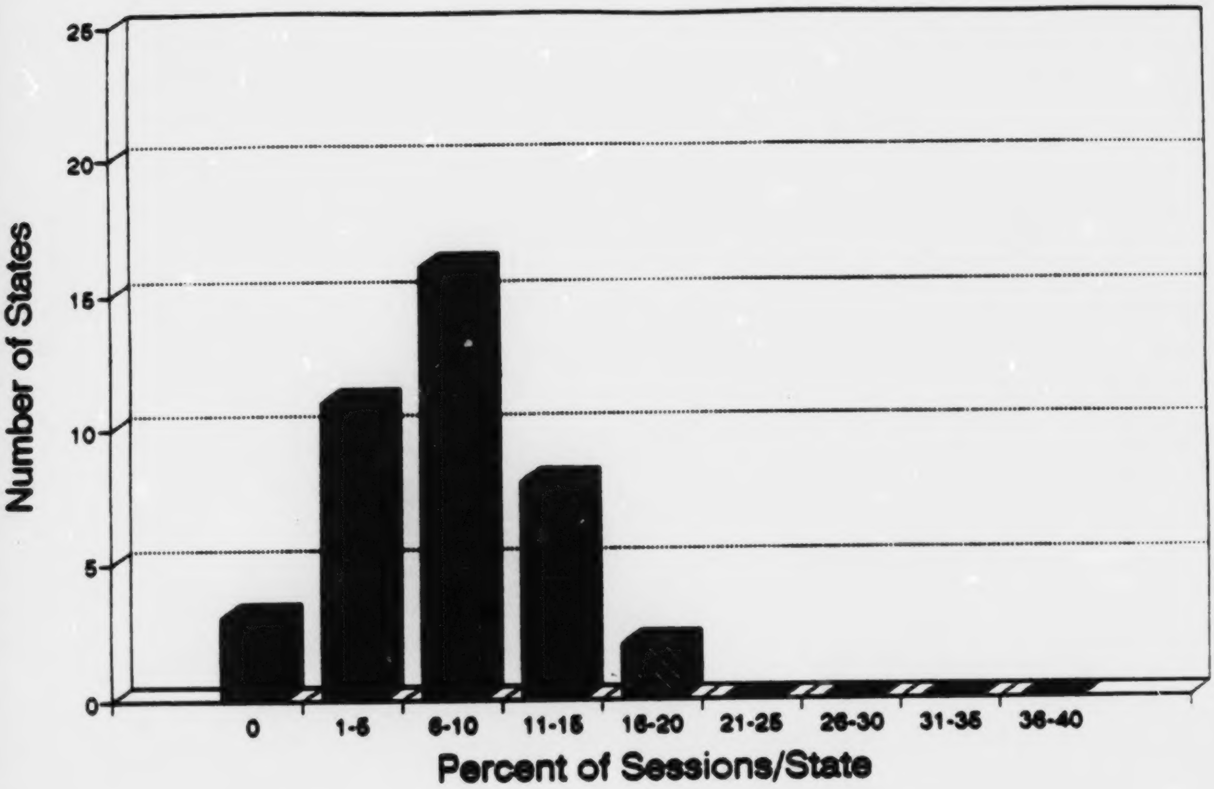


Figure 10. Student Disruption



Lower average scores were observed in sessions in which student disruption occurred, as shown in Table 30. The mean adjusted performance score for sessions with student disruption was -0.091 compared to 0.018 for the remainder, t = -3.13 p ≤ .01. Similarly, in sessions where student attitude was evaluated as "mixed/indifferent" or "negative," the mean adjusted performance score was -0.067 compared to 0.039 for the remainder, t = -4.25 p ≤ .001. This relationship is shown in Table 31.

Table 30. Means and Variances of Adjusted Performance Scores Problems with Student Disruption (N = 1,918 Sessions)

	No. of Sessions	Mean	Within-Session Variance	Between-Session Variance	Between-state Variance
With Problem	136	-.091**	.766	.071	.036
Without Problem	1782	.018	.765	.076	.031

** Significant at p ≤ .01

Table 31. Means and Variances of Adjusted Performance Scores Student Attitude Indifferent or Worse (N = 1,918 Sessions)

	No. of Sessions	Mean	Within-Session Variance	Between-Session Variance	Between-state Variance
With Problem	543	-.067***	.760	.083	.039
Without Problem	1375	.041	.768	.070	.028

*** Significant at p ≤ .001

The correlation between session mean measures of student performance and student attitude was significant (r = 0.274 p ≤ .0001). One might expect that both performance and attitude might be correlated with SES, but there was actually little or no correlation with either percent eligible for federal lunch program (0.043, ns, for student disruption, and 0.069, $p \le .0038$, for student attitude "mixed/indifferent" or worse) or mother's education (0.008, ns, for student disruption, and -0.37, ns, for student attitude).

Multivariate analysis. In an effort to assess the overall impact of administration irregularities on student performance, variables representing one or more errors in timing test sections, two or more errors in calculator training, one or more errors in calculator distribution, and three or more errors in script reading were entered into a multiple regression equation along with a number of control variables (e.g., mother's level of education, percent of students eligible for federal lunch program, student's perceived attitude towards the assessment, student disruption during the administration, and overall rating of LA mixed or worse). Not surprisingly, mother's education was the strongest predictor of student performance, even given that the measure of student performance had already been adjusted for effects of ethnicity and sex. Mother's education was followed by student attitude, percent eligible for federal lunch program, and student disruption. However, when all of these conditions were controlled, errors in calculator training and script reading still had a small but significant impact on student performance. (On the other hand, neither the influences of timing or calculator distribution errors nor the overall rating of LA performance was statistically significant.) These results, which are portrayed in Table 32, strongly suggest that, other factors being equal, a lack of clarity in the directions provided by the LA *does* depress student performance.

Table 32. Multiple Regression Predicting Mean Adjusted Performance Scores (Adj $R^2 = .403$)

Predictor	Standardized Parameter Estimate	T for HO: Parameter=0	Prob > |T|
Mother's Education	0.569	28.17	0.0001
Pct Eligible for Federal Lunch Program	-0.093	-4.61	0.0001
Student Attitude Indifferent or Worse	-0.095	-4.86	0.0001
Student Disruption	-0.057	-2.93	0.0034
Overall LA Rating Mixed or Worse	0.012	0.59	0.5542
^ 1 Errors in Timing Sections	-0.025	-1.31	0.1898
^ 2 Errors in Calculator Training	-0.044	-2.32	0.0203
^ 1 Errors in Calculator Distribution	0.006	0.33	0.7454
^ 3 Errors in Script Reading	-0.038	-2.04	0.0420

AIR staff also observed three trial state NAEP administrations in northern California. These observations generally confirmed the results of the QCM monitoring in that these administrations were observed to go smoothly with no major problems. Minor problems included some confusion with calculator distribution and the use of the timers, and one LA hinted that a student may have been suspended to keep him from participating in the assessment.

Validity Issues Related to Incentives and Student Preparation

Level of effort can profoundly influence test results. In a national assessment of this kind, where the state, and not the student or even the school, is the unit of analysis, the student's natural motivation to do well in order to secure a superior grade is lacking. It may even be difficult to call upon motivation based on a desire to uphold the reputation of the school. Some students may still be motivated by a more general sense of regional or state pride or a desire to meet their own internal standards, but such motivations are probably distributed unevenly in the population and very likely correlated with overall academic achievement. Further variation in motivation may be introduced if some school officials independently decide to offer their students incentives for participation or provide special information or instruction designed to make the test process more salient for students. (Extracurricular instruction aimed at improving mathematics skills or general test-taking skills could *directly* influence the student's performance as well as his/her motivation.) Interventions designed to enhance the student's concern for test outcomes would presumably have a positive effect on motivation and hence level of effort. Interventions designed merely to ensure student *participation* might have either a negative or positive effect. (As a related point, some states offered incentives to *schools* in order to obtain their participation. Such incentives may have either positively or negatively influenced the level of effort put forth by local school personnel in assuring the integrity of the TSA.)

There was concern that in some schools students may have received special instruction in the use of calculators prior to the session and that as a result, their scores would not be representative of the scores of typical students in the school. In 86 of 1,918 monitored sessions, such instruction was reported; yet as shown in Table 33, scores were generally lower in those sessions. Unless one were to conclude that the special instruction was actually harmful, this suggests that the special instruction was given in schools in which the need for it was greatest: these students might have scored even lower had the special instruction not been given. This explanation is plausible if it is supposed that students in many schools regularly receive instruction in the use of calculators while students in others do not. Since the scores in Table 33 have been adjusted for ethnicity and sex differences, the differences cannot be due to a correlation between ethnicity or sex of the students and the occurrence of special calculator instruction.

Table 33. Means and Variances of Adjusted Performance Scores for Sessions Before which Special Calculator Instruction was Given

Standardized Logit Scores	Sessions with Special Calculator Instruction	Sessions without Special Calculator Instruction
Mean	-.142	.018
Within-Session Variance	.685	.769
Between-Session Variance	.085	.075
Between-State Variance	.058	.029
t-value for mean difference	-3.08*	

* Difference significant, $p \leq .01$.

There was also concern that pep talks or meetings with the students prior to the testing session might artificially raise scores. In 297 monitored sessions prior pep talks were reported, and in 447 monitored sessions prior meetings were reported. However, as can be seen in Tables 34 and 35, the scores in these sessions were not significantly higher than scores in other sessions. Clearly, although racial-ethnic and sex differences are eliminated from these comparisons, further analyses, taking demographic factors into account, are necessary before concluding that such activities have no effect on performance.

In addition, there was concern that giving students specific incentives might alter their performance levels. As shown in Table 36, scores were, if anything, lower in sessions in which student incentives were reported to have been given.

Table 34. Means and Variances of Adjusted Performance Scores for Sessions Before which a Pep Talk Was Given

Standardized Logit Scores	Sessions with Pep Talks	Sessions without Pep Talks
Mean	-.004	.013
Within-Session Variance	.752	.768
Between-Session Variance	.077	.076
Between-State Variance	.020	.034
t-value for mean difference	-0.50	

Table 35. Means and Variances of Adjusted Performance Scores for Sessions Before which a Meeting Was Held

Standardized Logit Scores	Sessions with Meetings	Sessions without Meetings
Mean	.024	.006
Within-Session Variance	.778	.762
Between-Session Variance	.090	.071
Between-State Variance	.018	.036
t-value for mean difference	0.58	

Table 36. Means and Variances of Adjusted Performance Scores for Sessions in which Students Received Incentives

Standardized Logit Scores	Sessions with Incentives	Sessions without Incentives
Mean	-.041	.015
Within-Session Variance	.728	.769
Between-Session Variance	.046	.079
Between-State Variance	.037	.031
t-value for mean difference	-1.35	

Finally, the position of the LA within the local district may have had an effect on the seriousness with which students approached the assessment session. To test this hypothesis, scores in the 400 monitored sessions conducted by principals, vice principals, and district superintendents were compared with scores in sessions administered by others (primarily counselors and teachers). As shown in Table 37, there was a significant correlation between scores and the organizational level of the LA. Unless this is due to other correlates of the LA selection (e.g., size of school), it suggests that factors that increase students' perceptions of the seriousness of the assessment may increase their motivation, leading to higher scores.

Table 37. Means and Variances of Adjusted Performance Scores as a Function of the Organizational Level of the Test Administrator

Standardized Logit Scores	Sessions Administered by Principals, Vice Principals, and Superintendents	Sessions Administered by Counselors, Teachers, and Others
Mean	.080	-.008
Within-Session Variance	.751	.769
Between-Session Variance	.072	.076
Between-State Variance	.020	.034
t-value for mean difference	3.06*	

* Difference significant, $p \leq .01$.

Validity Issues Related to Creation of Data Files and Assignment of Student Characteristics

One responsibility of the LA was to maintain "administration schedules"--paper documents that listed each student in the original or supplemental sample and indicated the assessment outcome for that student (assessed, absent, withdrawn, or excluded). In addition, the LA recorded the student's birthdate, sex, and race, noted whether they were designated as IEP or LEP, and entered a code that identified the student's mathematics teacher and class period. A sample of administration schedules was obtained by AIR, and schedules for 25 sessions (covering 22 schools) were compared against computer files prepared by Westat and ETS. With regard to the total numbers assessed, absent, and excluded, there was perfect agreement between the hard copy schedules and a Westat computer file of assessment counts prepared directly from these schedules. Any arithmetic errors made by the LAs in summing the entries on the administration schedule had been appropriately corrected before the session totals were entered on the computer file.

LAs also were responsible for filling in answer grids for much of this same information on the covers of the assessment booklets used by the students. A comparison of the hard copy administration schedules with the data read from the booklet covers suggest that the LAs made relatively frequent errors in gridding. For example, one or more inconsistencies between the written and gridded information for math teacher ID number, math period, or both were observed in 10 of the 25 sessions examined. Errors included transposing the teacher and period numbers, apparently slipping one space on the grid, and leaving the grid information blank; frequent discrepancies were also observed in the data element for student year of birth. In all of these cases the values entered from the booklet covers were off by one year from the values written on the administration schedule. Types and numbers of errors are summarized in Table 38.

Table 38. Apparent Local Administrator Errors in Gridding Student Information onto Booklet Covers Sample of 25 Sessions in 22 Schools (Average of 24 Students per Session)

Data Element	# Errors	# Sessions w/Errors
Teacher ID number	14	8
Math period	10	7
Student birth year	14	9
Student birth month	2	2
Student sex	3	2
Student race (observed)	6	4
Student IEP or LEP status	0	0

These high error rates indicate that additional checks should be added to the system in future assessments to ensure the accuracy of field coding. In addition, the observed errors in teacher ID codes raise a cautionary note for any analyses linking information on teachers to student achievement data.

It is also noteworthy that, as might be expected, there were large numbers of discrepancies between the derived race values placed on the analysis files and the observed race values recorded by the LAs. The derived race values were based on student responses to two questions on race and Hispanic background. Observed race was used only if the student responses were missing or out of range. In the sample of 25 sessions, there were 34 instances, in 16 different sessions, where the derived race value was different from the observed race value recorded on the administration schedule. These included 20 instances in which the derived race was Hispanic and 14 other instances in which the derived race was black, Asian, or Native American. The primary implication of these latter findings is that any effort to evaluate the racial-ethnic distributions of the absent or withdrawn students (available only from the administration schedules) will be subject to a relatively high degree of error.

The administration schedules were also checked for discrepancies against the information coded from the covers of the excluded student questionnaires. For the 47 excluded students in the 25 sampled sessions, there was one student with no questionnaire record; one student whose record was missing information on birth month, birth year, and sex (although all of these fields were filled in on the administration schedule); and one student who had a different birth month on the computer file than on the administration schedule. There were also seven students who had race information on the computer file but not on the administration schedule. Since each excluded student questionnaire was completed by a teacher or counselor familiar with that particular student, discrepancies between the questionnaire and the administration schedule are probably appropriately resolved in favor of the questionnaire.

Feasibility Evaluation

One of the issues raised in assessing the future of the state NAEP revolves around the cost of implementation. The magnitude of the costs to state and LAs of implementing the state assessment will be a significant factor in their willingness to participate. Furthermore, to the extent that sources of variation in costs between states can be identified, recommendations for procedures to improve efficiency may be generated.

AIR carried out a survey of state testing directors, and in that survey 24 respondents indicated that the cost of the TSA was "about what we expected," while 6 found it "more expensive" and 4 found it "less expensive." Twenty-six respondents felt that the 1990 trial was likely to be worth the cost to their states, while 9 thought it was not likely. Nevertheless, because the tasks of the assessment are carried out by many individuals--state-level staff, LAs, local teachers, local guidance counselors, and many support staff employed in the school systems in a state--it is very difficult for states to estimate exactly the cost of participating in the assessment. Global estimates by the state testing directors for state-level costs ranged from $.96 per student to $31.79 per student, and estimates of local-level costs ranged from $.21 per student to $24.02 per student (based on an average of 2,500 students assessed per state).

Do cost estimates affect states' decisions to participate in the state assessments? Based on the self-reported cost data in the AIR survey, those states reporting that the costs were "more expensive" than expected reported the highest costs, while those states reporting the cost was "less expensive" than expected reported lower than average costs. Moreover, states responding that the 1990 trial was "not likely to be worth the cost" also tended to report higher costs than those responding that the trial was "probably worth the cost."

The extremely broad range of estimates is disturbing, even more so because there was virtually no correlation between cost estimates gathered by AIR and cost estimates gathered a few months earlier by Westat. This result suggests a need for a framework for estimating test administration costs. To facilitate cost estimation for future participation and the estimation of costs for different scenarios, AIR developed a resource utilization model for the various tasks involved in the TSA and conducted a detailed survey of seven state coordinators regarding allocations of personnel and nonpersonnel resources to state and local activities related to the implementation of TSA.

AIR selected eight state coordinators who had indicated that they were "very" or "quite" certain in their estimates of state-level costs of the trial assessment. A framework was developed for collecting detailed data on the allocations of personnel time and nonpersonnel resources to specific activities involved in the state coordination and local test administration. The activities specified are based on the responsibilities outlined for the state coordinator and the LA. Each of these components is described below.

Responsibilities and Activities of the State Coordinator

The tasks required at the state-level for participation in TSA are:

- attending LA training sessions,
- attending other related meetings (e.g., ETS' network meetings),
- getting the 100 schools to participate in the TSA program,
- identifying individuals at the local level to serve as local administrator for NAEP,
- scheduling dates for test administration,
- obtaining lists of potential eigth graders from the participating schools, and
- observing test sessions, as well as carrying out other miscellaneous tasks.

In addition to these direct costs, administrative overhead ratios should be applied to the hourly costs of the state coordinator to cover the costs of office facilities, utilities, equipment, and the costs of administrative supervision of the state coordinator.

Responsibilities and Activities of the Local Administrator

The costs of local test administration result from the activities carried out by the LA. Each of these activities is listed below. Unless otherwise indicated, these activities involve primarily the time of the LA.

- submitting list of all eigth grade students,
- reviewing the LA's manual,
- attending (5-hour) training session,
- receiving the list of students to be tested,
- scheduling a room for testing,
- ensuring the completion of questionnaires by math teachers, counselors, and principals,
- administering test sessions,
- preparing and shipping test materials, and
- conducting other activities (e.g., make-up sessions if needed).

Two data collection forms were designed to gather the allocations of personnel and nonpersonnel resources to these activities. These forms are presented in Appendix B. Copies of these forms were provided to the state coordinators in 8 states to acquire more detailed information on allocation of resources to the implementation of the Trial State Assessment. One of the eight states provided only aggregate data on time estimates and was excluded from the analysis. One of the remaining seven states did not provide any data on local test administration.

Any conclusions based on this resource-cost study must be tentative. The data collection was carried out nearly a year after the implementation of the assessment, and reconstruction of some of the time requirements of these activities as problematic for some respondents. Furthermore, the state coordinator was the respondent for both state and local cost estimates. While this might be a legitimate way of gathering data for the state costs, it is likely that more accurate estimates would be obtained by collecting the local resource/cost data directly from LAs. In either case, those individuals responsible for developing the database for evaluation of the TSA will need to plan for systematic data collection directed at both the state coordinators and the LAs. With advance notice to these parties that questions regarding levels of effort and time allocations will be asked, these data can be tracked relatively easily and reported with significantly greater accuracy for the purposes of developing and analyzing cost estimates.

Based on the time estimates reported in the selected subsample of states, Table 39 presents the range of costs for state- and local-level activities. State-level costs ranged from $1.25 to $6.65 per student, while local-level costs ranged from $16.86 to $85.28 per student. These figures clearly reflect a wide range of costs, and one can now explore the factors underlying these variations. State #4, which reported the highest per-student costs for local activities, allocated a considerable amount of secretarial and clerical time at the local level in contrast to the other four states where virtually no secretary and clerical time was allocated. In addition, the costs of time for teachers, counselors, and principals to complete the student and school questionnaires was explicitly recognized in the state #4 estimates, while only one of the other four states (i.e., state #5) reported any time for this activity.

Table 39.	**Costs of Trial State Assessment as Reported on Detailed Telephone Survey of State Coordinators***	
	Cost per Student for NAEP Activities	
	State-Level Costs	Local-Level Costs
State #1	$1.25	$16.86
State #2	$4.18	$21.06
State #3	$2.76	No data available
State #4	$6.65	$85.28
State #5	$4.73	$47.59
State #6	$3.76	$17.49
State #7	$3.43	$27.89

[*] It is important to point out that the costs reported in Table 39 do not reflect any overhead costs associated with office space and equipment, utilities, or administrative supervision of the individuals involved in test administration. Given the nature of the organizational arrangements for conduct of the Trial State Assessment, this is likely to affect only the state-level costs since the supervision of the LAs is assumed by the NAEP subcontractor (Westat). Although no information on overhead costs was collected, experience suggests that it would not be unreasonable to add another 25 percent to the state-level costs to account for overhead.

The process utilized for gathering the detailed cost data from state coordinators revealed to some of these individuals that they had not considered all of their costs in their original estimates reported in the earlier survey of state testing directors. For example, one state coordinator reported that his state had not included the time of any personnel in the original cost estimates, but rather had considered only out-of-pocket costs for non-personnel resources (e.g.,travel). Other states indicated that the estimates derived from the detailed questionnaires provided a more precise estimate of the costs than those reported on the original forms. In two of the cases, the cost estimates were obviously derived from a process similar to the one reflected in the detailed questionnaires and matched fairly closely. In fact, for the six overlapping observations between the earlier AIR survey and the detailed cost data reported in Table 39, there is a positive, but not significant, correlation of .276. This positive correlation at least suggests some similarity in the basis for the two cost estimates, although the correlation is not strong. A negative correlation of -.383 exists between the earlier Westat time and cost estimates and the detailed cost data reported in Table 39, suggesting that the data underlying these cost estimates are not the same.

Prototype estimates of costs. In order to address the feasibility of NAEP state assessments, one goal of this study was to develop reasonable estimates of what it actually costs to implement the TSA. The proposed structure for cost analysis, combined with the data gathered from the seven states, provide a foundation for prototype cost estimates of state- and local-level activities related to implementation of the TSA. Drawing on these data and on information gathered through visits to local sites to observe the TSA implementation process, AIR has developed a reasonable range of time estimates for the conduct of the various state- and local-level activities in the 1990 TSA. These time estimates are presented in Tables 40 and 41. These prototypes were based on a detailed evaluation of the responses given for each category of state and local task involved in the TSA. Table 40 presents the prototype time and cost estimates for the state-level activities, and Table 41 presents the time and cost estimates for the local-level activities. In each case, one low range and one high range time and cost estimate are presented. This range of estimates represents reasonable upper and lower boundaries for the actual costs of implementing the TSA.

Table 40. Prototype Time Estimates for State-Level Activities

Low Estimate

Activity	Hrs of pers. time		Travel
	State coord.	Clrk/ secr.	
Attend training sessions	8.00	2.00	
Attend other meetings	53.00	2.00	
Get 100 schools	1.00	4.00	
Identify local administrators	1.00	1.00	
Set testing dates	2.00	2.00	
Schedule training	2.00	8.00	
Obtain lists of eighth graders	4.00	4.00	
Conduct other activities	40.00	2.00	$125
Total hours	111.00	25.00	
Hourly wage rate	$30.00	$15.00	
Hourly benefit rate	$ 9.00	$ 4.50	
Total cost by resource	$4,329	$488	$125
Total cost--state-level	$4,942		
Number of students tested	2,500		
Cost per student	$1.98		

Table 40. Prototype Time Estimates for State-Level Activities (cont.)

High Estimate

Activity	Hrs of pers. time		Travel
	State coord.	Clrk/ secr.	
Attend training sessions	16.00	2.00	
Attend other meetings	106.00	6.00	
Get 100 schools	10.00	24.00	
Identify local administrators	10.00	5.00	
Set testing dates	2.00	8.00	
Schedule training	4.00	8.00	
Obtain lists of eighth graders	10.00	10.00	
Conduct other activities	90.00	10.00	$250
Total hours	248.00	73.00	
Hourly wage rate	$30.00	$15.00	
Hourly benefit rate	$ 9.00	$ 4.50	
Total cost by resource	$9,672	$1,424	$250
Total cost--state-level	$11,346		
Number of students tested	2,500		
Cost per student	$4.54		

Table 41. Prototype Time Estimates for Local-Level Activities

Low Estimate

Activity	Hrs of pers. time					Travel
	Test admi.	Clrk/ Secr.	Tchr	Cnslr.	Princ.	
Submit list of eighth graders	1.00	2.00				
Review LA manual	2.00					
Attend training session	8.00					$100
Receive list of students to test	1.00					
Schedule room for testing	0.25	0.25				
Ensure questionnaires completion	0.50		0.75	0.75	1.25	
Administer test sessions	3.00					
Prepare/ship test materials	1.00					
Carry out other tasks					1.00	
Total hours	16.75	2.25	0.75	0.75	2.25	
Hourly wage rate	$25.00	$15.00	$25.00	$27.00	$35.00	
Hourly benefit rate	$ 7.50	$ 4.50	$ 7.50	$ 8.10	$10.50	
Total cost by resource	$544	$44	$24	$26	$102	$100
Total cost--per session	$840					
Number of students tested	25.48					
Cost per student	$32.97					

Table 41. Prototype Time Estimates for Local-Level Activities (cont.)

High estimate

Activity	Hrs of pers. time					Travel
	Test admi.	Clrk/ Secr.	Tchr	Cnstl.	Princ.	
Submit list of eighth graders	7.00	4.00				
Review LA manual	4.00					
Attend training session	16.00					$125
Receive list of students to test	3.00					
Schedule room for testing	1.00	1.00				
Ensure questionnaire completion	1.00		4.50	1.50	1.50	
Administer test sessions	4.00					
Prepare/ship test materials	1.50					
Carry out other tasks	2.00	3.00	2.00		2.00	
Total hours	39.50	8.00	6.50	1.50	3.50	
Hourly wage rate	$25.00	$15.00	$25.00	$27.00	$35.00	
Hourly benefit rate	$ 7.50	$ 4.50	$ 7.50	$ 8.10	$10.50	
Total cost by resource	$1,284	$156	$211	$53	$159	$125
Total cost--per session	1,988					
Number of students tested	25.48					
Cost per student	$78.03					

No variations in salaries were included in the prototype estimates since the focus of this analysis is to examine the variations in the levels of effort required to implement the TSA. While different regions of the country will exhibit different patterns of compensation for employees, these differences in compensation are a result of factors that are beyond the control of those individuals making decisions about the conduct of the Trial State Assessment. Such differences are a result of variations in the conditions and structure of local labor markets. Therefore, it is appropriate to remove the impact of variations in wage rates from prototype cost estimates in order to isolate the effects on costs of differences in the time allocated to different activities. Obviously, a particular state desiring to use the prototype to estimate its own costs would use hourly wage and benefit estimates that reflect its actual wage and benefit rates.

Each of the elements that went into the construction of the prototype estimates is discussed in some detail below. The discussion is organized around the various activities listed on the cost/time estimate worksheets utilized for the analysis.

Costs of State Coordination

Attend LA training sessions. The attendance of the state coordinator at the LA training sessions was not mandatory; however, state coordinators were encouraged to attend at least one session. Therefore, a low of 8 hours up to a high of 16 hours to allow for the possibility of attending more than one session was estimated, including some time for travel. Two hours of clerical time was allocated for making travel and attendance arrangements.

Attend other meetings. State coordinators attend up to six network meetings per year. Based on casual observation, it appears that about two thirds of the state coordinators attend these meetings, on the average. At the low end 12 hours per meeting was allocated for states close to the meeting site (Washington, DC), while at the high end three full days of time was allocated (i.e., allowing a day of travel each way). NCES pays all of the travel costs for these meetings, so travel costs are not included in the estimates. An additional 5 to 10 hours was added to the low and high estimates respectively to account for in-house meetings with staff and Westat staff coordinating the TSA activities.

Get 100 schools to participate. This activity involves drafting a letter of request for participation (approximately one hour) and using an assistant to address and mail the letter to 100 principals (approximately four hours). For states mandating participation, this may be all that is necessary. For states not mandating participation, additional time is required to call and encourage recalcitrant schools to participate. These kinds of contacts involve additional time from a secretary and the state coordinator. A low of 1 and 4 hours for the state coordinator and the assistant and a high of 10 and 24 hours for each was estimated.

Identify individuals to be LAs. This task is done for the most part at the local level, so it should be included on the second worksheet under other activities. Only a small amount of time was allocated in this worksheet to reflect a monitoring function on the part of the state coordinator and the clerical assistant. Nevertheless, the finding that students scored higher in sessions administered by principals, vice principals, and superintendents should be considered in weighing the costs and benefits of who is asked to serve in the role of LA. At the very least, the fact that statistically significant differences do appear suggests that LAs should be selected from among employees whose job titles are not associated with any significant differences, so the results are at least consistent across states.

Set dates for test administration. States are divided into quadrants; testing takes place for one week in each quadrant. Presumably, the state coordinator would have an assistant call the approximately 25 schools in each quadrant and decide upon the times. This might take anywhere from two to eight hours, depending on how many callbacks were required to make contact.

Schedule training sessions. This process is somewhat more time-consuming than setting dates for test administration because the sessions must be scheduled to minimize travel time for LAs. A range of from two to four hours of the state coordinator's time and about eight hours of the clerical assistant were estimated.

Obtain lists of potential eighth graders from 100 participating schools. This job requires oversight by the state coordinator, as well as substantial clerical assistance. It involves contacting the school administration office (by phone or by letter), requesting the information, and ensuring that the information received is what was requested. It was estimated that a low of 4 hours each for the state coordinator and the clerical assistant and a high of 10 hours each, respectively, were required for this activity.

Conduct other activities. Under other activities, the prototype includes an estimate of the amount of time to attend 5 to 10 test sessions at about 8 hours per session plus from 0 to 10 hours of time for other unspecified activities. Additional clerical time is included to assist the state coordinator and help make travel arrangements. Travel costs were included under the assumption that only local travel is required (with no overnight stay).

Costs of Test Administration at Local Sites

Submit list of all eighth grade students. Depending on how well organized the school or district data system is, this task could range from fairly straightforward to relatively tedious. The prototype estimates range of from one to seven hours for the local testing administrator and from two to four hours for a clerical assistant.

Review LA manual. It was estimated that the task of reviewing the testing manual would require from two to four hours.

Attend training session. The training sessions themselves are scheduled to last about five hours. Allowing some time for travel, it was estimated that attendance plus travel would range from 8 hours to 16 hours, depending on the extent of required travel time (e.g., in large states). Since training sessions were held in different parts of the states for the convenience of the local testing administrators, travel costs would range from $100 to $125. (As was described earlier in the report, some of the LAs indicated that the training sessions were excessive. However, AIR found that deviations from the uniform procedures that occurred were generally correlated with slightly lower mean scores, so the extent of training should probably not be reduced.)

Receive and process list of students to be tested. The LA must go through the list, update it, and copy information about student ethnicity, sex and classroom at this point. In addition, the LA must identify the appropriate math teachers and the students excluded from testing. For a list of 30 students, this task would take from one to three hours depending on the accessibility of the information.

Schedule a room for testing. Depending on the availability of space and the processes for reserving the space, this task could take from 15 minutes to one hour for the local testing administrator.

Completion of questionnaires. The principal completes the principal questionnaire and a school policies questionnaire, with the two requiring about 1.5 hours total. (An alternative would be to have the LA or a counselor complete the school policies questionnaire.) The excluded student questionnaire is completed by a counselor and requires no more than 30 minutes per student. It is estimated that about 1.5 students per school were excluded, requiring an estimated time allocation of about 45 minutes. The mathematics teacher questionnaires included requests for information about the teacher's background as well as about the class in which each student takes mathematics. In a sample of 10 sessions, there were anywhere from one to seven teachers and 2 to 13 different classrooms represented among the 25 to 30 students tested. The teacher background questionnaire is estimated to take about 30 minutes and the brief column of information on each classroom is assumed to take about 5 minutes. Rounding to the nearest quarter of an hour, the time estimates for completing the mathematics teacher questionnaires range from about three-quarters of an hour to approximately 4.5 hours.

Administer test session. The test session itself usually takes about two hours with about one to two hours' preparation on the part of the LA. In approximately 1 percent of the cases, this was doubled because of the need for make-up sessions.

Prepare and ship test materials. Preparation of the materials for shipping requires anywhere from one to one and half hours of the time of the LA.

Carry out other tasks. Some additional time was allocated for the LA (two hours), secretary (three hours), and teachers (two hours) for the high estimate to account for otherwise unspecified activities. An additional hour of the principal's was also allocated to identify and enlist the time and effort of the local testing administrator at the school site.

Table 42 summarizes the results of the prototype time estimates presented in Tables 40 and 41. These prototype estimates of state coordination activities range in cost from about $1.98 per student to $4.54. For local test administration the estimates range from $32.97 to $78.03. Combining these two estimates leads to a range of $34.95 to $82.57 per student. Although these still represent a fairly wide range of costs, the estimates do narrow the range of total costs somewhat from the data derived from the questionnaires presented earlier. These estimates provide a benchmark for states to use in developing estimates of the total costs of participating in the TSA. The estimates also provide a point of departure for improving the quality of the estimates and some information on which to evaluate alternative approaches for reducing costs.

Using the overall average of the prototype figures, the cost per student might be reasonably estimated at about $58.76. If approximately 100,000 students were included in the testing, the overall costs of the Trial State Assessment range from $3.5 to $8.3 million with an average of about $5.9 million. If each state were to test approximately 2,500 students, the overall costs of the state's participation in TSA would range from $90,000 to $210,000, with an average of about $150,000.

Comparatively, one can easily see that the costs of state coordination probably do not offer much opportunity for savings because of the relatively small part of total cost that they represent. If any major savings in the costs of implementation are to occur, they will likely be at the local level. Focusing for a moment on the local component, one can see that the largest single activity occurs in training the LA (8 to 16 hours in the low and high estimates). If one LA could be utilized across multiple testing sessions, these costs of training, as well as the costs of reviewing the instruction manual, would be spread over a larger number of students. This strategy could include multiple sessions at a given school or across more than one school. One would have to balance additional travel time required for one LA to cover multiple schools against the time saved by having to train fewer individuals for this responsibility. Having a single LA cover multiple sites and sessions is also likely to create some potential economies in handling testing materials. On the other hand, each school may perceive an individual benefit by having a NAEP-trained LA.

Table 42. Summary of the Costs of the Prototype

	Low Estimate	High Estimate
Cost of state coordination		
Total Cost	$4,942	$11,346
Number of Students Tested	2,500	2,500
Cost per Student	$1.98	$4.54
Costs of local administration		
Total Cost per Session	$ 840	$ 1,988
Average No of Students per Session	25.48	25.48
Cost per Student	$32.97	$78.03
Total cost per student	$34.95	$82.57

Differences in wage rates can be achieved in the choice of the level of employees assigned as state coordinator and LAs. Focusing again on the local level, it is reported that counselors were utilized as LAs in about half of the sites. In about 17 percent to 20 percent of the sites, either principals or vice principals were used as LAs. It is reasonable to assume an average hourly wage of about $25 plus a 30 percent benefit rate for the LA in generating cost estimates.

With the serious discussion that has occurred surrounding the possibility of implementing a national examination, there is likely to be significant value in improving the quality of these time and cost estimates. With a strong analytical database on the costs of implementing TSA, one can begin to conduct analyses of alternative approaches to administering tests to students that would be useful in estimating the costs of a national examination. For example, with an analytical database on the cost of test administration, one could estimate the impact of increasing the number of school sites, the number of sessions per school, the number of students per session, and the type of data collected (e.g., portfolio assessment) on the costs of state- versus local-level activities. Each of these components would have relevance in designing the structure of a national examination.

Summary and Conclusions

The National Academy of Education, through their subcontractor, the American Institutes for Research, carried out an evaluation of the implementation of the 1990 Trial State Assessment in eighth grade mathematics. The evaluation involved observation of training and testing, an independent survey of state testing directors, and analyses of preliminary data files produced by the NAEP contractor, the Educational Testing Service (ETS), and its subcontractors, Westat, Inc. and National Computing Systems (NCS).

The evaluation focused on issues related to the validity of the test results obtained and on the feasibility of state-level assessment in NAEP in terms of costs from the local and state perspectives. Validity issues were addressed in four broad areas: (1) representativeness of the students actually assessed, in comparison to the target population, (2) transmission of materials to the testing sites and preparation for the test administration, (3) actual administration of the test session, and (4) the effects of variation in participant incentives and attitudes.

The government's contractor for NAEP, ETS, was well-prepared for the Trial State Assessment. Training sessions conducted by the NAEP data collection subcontractor, Westat, for both state coordinators and local administrators were observed to be highly professional, covering the assessment quite thoroughly. Contractor's quality control monitors (QCMs) observed half of the sessions and generally reported a high level of accurate compliance with the instructions for administering NAEP.

Most of the analyses reported here differed from those being carried out by ETS in two major ways. (1) Each assessed student was weighted equally because the goal of the study was primarily to analyze the sample results, not to generalize to state averages. (2) A simple transform of "percent correct," rather than a final scale score, was the dependent variable in comparisons. In addition, most of the comparisons were adjusted for differential race and sex compositions of school samples, so that results would not be confounded by race and sex effects. The major findings of the independent evaluation of the implementation are summarized below.

Were there differences between monitored versus unmonitored sessions?

Only very small test score differences were found between monitored and unmonitored sessions, in either mean levels or variances. However, in some sessions, monitors did intervene in order to preserve the validity of the administration. Furthermore, valuable data on the administration were collected from monitored sessions.

Recommendation 1: Continue to monitor a substantial proportion of the sessions.

Were there differences in scores between TSA and national NAEP?

The aggregate average TSA performance score was 0.12 standard deviations higher than the aggregate average comparison sample performance, on a measure with a nominal population standard deviation of 1.00.

A similar comparison of reported "mother's education" for the two samples showed nearly an exact match, indicating that the difference between the two samples was more likely to be related to conduct of the assessment than to representativeness of the assessed students.

These comparisons indicated, as a by-product, that a substantial proportion of between-state score variation is associated with differences in racial-ethnic distributions.

Did school refusal and substitution affect outcomes?

Approximately 6 percent of all originally invited schools refused to participate in the TSA, although in two states the proportion approached 20 percent. The main differences between cooperating and refusing schools, which were statistically significant in some states, were that smaller schools tended to refuse more frequently and that schools with lower percentages of black students tended to refuse more frequently.

Although the mean scores in substitute schools were slightly higher than scores in originally cooperating schools, that may merely reflect that the substitute schools matched the refusing schools in being smaller and having fewer black students than cooperating schools, on the average.

Did "withdrawnstudents" and "absent students" affect outcomes?

It is impossible to know how withdrawn or absent students would have scored, but evidence was found relating numbers of withdrawn or absent students to scores of those present for the assessment. There was a significant *negative* correlation, -0.27, between the number of students withdrawn from a session and the average performance score of those present in that session. Virtually the same negative correlation was observed between numbers of absent students and average scores. The most straightforward interpretation of these results is that schools with high rates of student absences and departures are also schools with lower mathematics achieving students.

One state differed from all others in the frequency of absences, and the presentation of results for that state should be qualified by the large absence rate.

How often and with what effects did pre-assessment problems occur?

Quality control monitors called each local administrator a few days before the assessment to find out whether materials had been received and whether questionnaires for mathematics teachers and counselors had been distributed. Although instances of non-receipt of materials were rare, these calls were probably necessary because in many cases the questionnaires had not been distributed. These calls undoubtedly increased the rate of return of these questionnaires, which are a rich source of information on the educational context of each assessed student. Average performance scores, on the other hand, were not significantly related to pre-assessment problems.

How often and with what effects did administration irregularities occur?

Timing errors

Timing errors occurred, but they were almost entirely in the section that called for the student's background information, which was the first timed section and had no bearing on performance scores. Nevertheless, average performance scores were lower in sessions in which timing errors were reported. The most straightforward interpretation of this is that the occurrence of timing errors is an indicator of some more basic factor, such as student motivational variation. In the future, recording the direction of timing errors (allowing too little or too much time) would increase the value of these data.

Recommendation 2: Add monitoring items to indicate whether each error in timing resulted in either more time or less time for a particular section.

Calculator distribution

No decrement in student performance was found to be associated with errors in calculator distribution, but these results might have been different if the QCMs had not intervened in virtually all instances of distribution error.

Script reading

Errors in reading the script were the most frequently observed variations in administration, but they tended to be distributed randomly over the participating states, with no state as a significant outlier. Scores tended to be lower in sessions in which deviations from the script, especially in reference to how to use the test booklet, were reported.

Recommendation 3: **Emphasize in LA instructions, even more than in the past, the need and the value of following the script in giving instruction.**

Did variation in incentives or motivation affect scores?

There is concern that test results might be affected by differential incentives to students. Generally, and somewhat surprisingly, average scores in sessions in which students were reported to have been given incentives, a pep talk, or special instructions were slightly lower than in others.

Sessions were conducted by counselors, teachers, principals, and other educators. Scores in sessions conducted by principals, vice principals, and superintendents were found to be higher than in other sessions. The most straightforward interpretation of this finding is that the apparent position of power of the test administrator may have affected students' perceptions of the importance of good performance on the test.

Recommendation 4: **To facilitate understanding of affective impact, add monitoring information on who was present in each session, and qualitatively, for how long.**

Some of the strongest associations with average session scores were found for the items that measured student attitudes towards the assessment and behaviors during the assessment. Sessions in which students had negative attitudes or displayed disruptive behavior were characterized by lower scores.

Recommendation 5: **Add behaviorally anchored ratings of student behavior to the LA questionnaires in both monitored and unmonitored sessions. The anchors might be:**

1. **Students were very hard to control, and some (at least 5) appeared not to try at all to answer the assessment questions.**

2. **Students were hard to control, but nearly everybody appeared to try to answer the assessment questions. At least one student behaved in a manner that would affect other students' performance.**

3. **Nearly all students were attentive, with lapses only during instructions or toward the ends of sections.**

4. **All students were attentive throughout the instructions and while answering the assessment questions.**

Recommendation 6: Identify whether or not script deviations by the LA were in response to unexpected events or behaviors in the room.

How expensive was the assessment, from the state perspective?

Most state testing directors perceived their costs generally to be in line with their expectations. Retrospective cost estimates varied widely among states because of the many ways in which components of the cost could be inadvertently overlooked (e.g., the time spent by mathematics teachers to fill out questionnaires). An initial analysis of costs by function suggested that the total costs for a state's participation were probably between $90,000 and $210,000, most of which were incurred at the local level. Most state testing directors felt that the assessment was worth the cost, and the number of participating states will apparently increase in 1992.

Recommendation 7: To facilitate accurate estimation of administration costs, add items on the amount of time required for various tasks (e.g., training, session administration, teacher questionnaire completion) to the LA debriefing information collected by the QCM.

Overall, the initial evaluation of the TSA indicates that the effort is probably feasible and probably provides the basis for reporting valid state-by-state results. A number of issues and concerns were raised, however, by the patterns observed in the data, and these should be addressed. First, monitoring of sessions provided valuable data and probably avoided many problems, and it should be continued. Second, issues such as the differences between the Trial State Assessment scores and the National NAEP scores need further investigation, especially as they relate to student motivation during the assessment. Finally, training of LAs should emphasize, even more, the value of uniformity of administration procedures. Although the administration was of high quality, the infrequent variations that did occur were often associated with lower average performance scores.

APPENDIX A

VARIANCE ESTIMATION AND SIGNIFICANCE TESTING

In conducting analyses of the administration data from the 1990 TSA, AIR focused on the file of 100,843 assessed students. Although ETS has constructed weights to represent the population, there was no deliberate oversampling in the TSA, so that treating each student as having the same weight would not introduce gross confounding. Therefore, since ETS had standardized the logit measure without using weights and since the purpose of the AIR study was to compare scores in groups of sessions that had different characteristics (e.g., occurrence or non-occurrence of timing errors), analyses were conducted weighting all assessed students equally.

In addition to reporting mean logit scores, AIR has reported variance components and levels of significance based on Student's t test. The purpose of this appendix is to document the specific computations carried out. The basic model for a score x_{ijk} is that it is the sum of an overall mean (μ), a state effect (α_i), a session effect ($\beta_{j(i)}$), and a student effect ($\varepsilon_{k(ij)}$):

$$x_{ijk}=\mu+\alpha_i+\beta_{j(i)}+e_{k(ij)}$$

The variance components to be estimated are:

$$\sigma^2_\alpha, \sigma^2_\beta, \text{ and } \sigma^2_\epsilon.$$

These are obtained by solving the three equations:

$$SS_1=(n_{..}-g_.)\,\sigma^2_\epsilon$$

$$SS_2=(n_{..}-snn_1)\,\sigma^2_\beta+(g_.-h)\,\sigma^2_\epsilon$$

$$SS_3=(n_{..}-snn_3)\,\sigma^2_\alpha+(snn_1-snn_2)\,\sigma^2_\beta+(h-1)\,\sigma^2_\epsilon,$$

where

SS_1 = Within-Session sum of squared deviations,
SS_2 = Between-Session Within-state sum of squared deviations,
SS_3 = Between-state sum of squared deviations,
n_{ij} = total number of assessed students in session j in state i,

g_i = total number of sessions in state i,
h = total number of states,

and

$$snn_1 = \sum_{i=1}^{h} \frac{\sum_{j=1}^{g_i} n_{ij}^2}{n_{i.}} \qquad snn_2 = \frac{\sum_{i=1}^{h}\sum_{j=1}^{g_i} n_{ij}^2}{n_{..}} \qquad snn_3 = \frac{\sum_{i=1}^{h} n_{i.}^2}{n_{..}}$$

When the solution included a negative estimate of a variance component, that entry was indicated by a "-"in the corresponding table cell.

Comparisons were carried out, for this report, between pairs of groups of sessions. The approximate Student's t was obtained by dividing the difference between means for the two groups, say *a* and *b*, by the square root of the estimate of the variance of that mean difference. The formula for the estimate of variance used was:

$$\sigma^2_{\bar{x}_a-\bar{x}_b} = \sigma^2_{\alpha}\left(\frac{snn_{3a}}{n_{..a}} + \frac{snn_{3b}}{n_{..b}} - 2\sum_{i=1}^{h} \frac{n_{i.a}n_{i.b}}{n_{..a}n_{..b}}\right) + \sigma^2_{\beta}\left(\frac{snn_{2a}}{n_{..a}} + \frac{snn_{2b}}{n_{..b}}\right) + \sigma^2_{e}\left(\frac{1}{n_{..a}} + \frac{1}{n_{..b}}\right)$$

where the variance components are averages of the two groups' estimates, weighted by the effective sample sizes of the groups. Because largest variance component was always the within-session variance, the estimated number of degrees of freedom for the test was always very large.

APPENDIX B
SURVEY INSTRUMENTS USED BY AIR

1990 TRIAL STATE NAEP EVALUATION
STATE TESTING DIRECTORS QUESTIONNAIRE
(for Participating States)

JULY, 1990

The following survey is being conducted by the American Institutes for Research, acting as a subcontractor to the National Academy of Education, as part of its evaluation of the 1990 Trial State National Assessment of Educational Progress. Please complete the following questions, answering as accurately, completely, and honestly as possible. Be assured that all individual responses will be held in strict confidence; only aggregate data will be reported. Thank you very much for your cooperation!

The questionnaire is divided into four parts, each containing roughly seven to ten questions. Part A deals with participation in the Trial State NAEP prior to February 1990; Part B contains questions about the actual administration in your state; Part C deals with policy issues regarding the usefulness of the State-by-State NAEP; and Part D asks for your opinions about the future of the State NAEP program.

PART A. The following questions refer to your participation in the 1990 NAEP Trial State Assessment (TSA) prior to the February 1990 data collection.

1. Do you feel that you were given adequate opportunity to provide input into the development of the 1990 Trial State Assessment? ____Yes ____No

Comments:____
____

2. Do you feel that you were kept adequately informed about decisions that were made prior to the administration of the 1990 Trial State Assessment? ____Yes ____No

Comments:____
____

3a. Did you receive the **State of Education** from ETS? ____Yes ____No
3b. Are you receiving the **NETWORK News** from ETS? ____Yes ____No

4. Do the NETWORK meetings provide an adequate forum to answer questions and address concerns you have about State NAEP? ____Yes ____No

Comments:____
____

5. Do you feel the need for another forum in addition to the NETWORK meetings to provide policy input from state testing directors? ____Yes ____No

If you answered "yes" to the preceding question, who should organize such a forum?

____ETS
____NCES
____NAGB
____CCSSO
____Some combination of ETS, NCES, NAGB and CCSSO, specifically:____________________
____
____Other
____Not applicable

6a. Who **ultimately** decided that your state should participate in the State NAEP? (please check one)

____State board of education
____State superintendent of education
____State legislature
____Governor's office
____Other

6b. Why did your state participate in the State NAEP? (please check all that apply)

____The state legislature mandated participation
____The Governor's office/superintendent's office/board wanted us to participate
____We felt that our state couldn't afford not to participate (because of the adverse publicity that would result)
____We participated voluntarily
____Other reasons (please specify):

____
____

7. What did your state hope to gain from participating in the State NAEP? (please check all that apply)

____My state is interested in the state ranking information
____My state is interested in obtaining school/district level data in the future
____My state is interested in international comparisons in the future
____We wanted to avoid the possibility of adverse publicity associated with non-participation
____We wanted a voice in shaping policy for the State NAEP
____We had no particular expectations for the 1990 State Assessment
____Other expectations (please specify):

____
____

= =

PART B. In this section, we would like to ask you some questions about conduct of the 1990 Trial State Assessment (TSA) in your state.

1. Procedurally, how did the 1990 Trial State NAEP administration go in your state?

____The administration went very well, there were no problems
____The administration went well, there were a few problems
____The administration did not go well

2. If there were problems with the administration in your state, what were they? (please check all that apply):

____Problems getting schools to participate
____Problems with the local administrator training
____Problems with the paperwork for school/student selection
____Problems with logistics at the school level
____Too much staff time was required for the administration
____Problems with receipt of testing materials from NAEP
____Communication problems with NAEP personnel
____Other problems (please specify):

____
____

3a. How much did it cost **at the state level** to administer the State NAEP?

Direct costs (actual dollars spent): ____
Indirect costs (non-reimbursed staff time): ____

3b. How certain are you about this estimate? ____Very ____Quite ____Not certain

3c. How much did it cost **at the local level** to administer the State NAEP?

Direct costs (actual dollars spent): ____
Indirect costs (non-reimbursed staff time): ____

3d. How certain are you about this estimate? ____Very ____Quite ____Not certain

4a. What is your view about the cost of administering the 1990 Trial State Assessment?

____The 1990 TSA was more expensive than we expected
____The cost of the TSA was about what we expected
____The cost of the TSA was less than we expected

4b. In your opinion, is the 1990 trial likely to be worth the cost to your state?

____Yes, definitely
____Yes, probably
____Probably not
____Definitely not

4c. If you believe the trial is probably or definitely not worth the cost, please briefly explain why:

____
____
____

5. What is your view about the amount of state, district and school staff time involved in the administration of the 1990 Trial State Assessment?

____It took more time than necessary
____It took about the right amount of staff time
____It took less staff time than it should have

6. What is your view about the amount of paperwork involved in the administration of the 1990 Trial State Assessment?

____There was more paperwork than necessary
____There was about the right amount of paperwork
____There was less paperwork than there should have been

7. What is your view about the amount of training involved in the administration of the 1990 Trial State Assessment?

____More training was given than was needed
____The amount of training was about right
____More training was needed

8. If you had to give an overall grade to the 1990 assessment in your state, what grade would you give?

____A ____B ____C ____D ____F

PART C. This section of the survey deals with policy issues regarding the utility of the state-by-state NAEP.

1. How will the results of the Trial State NAEP be used in your state? (please check all that apply):

____For publicity
____To advise curriculum development
____To argue for increased funding for education
____To inform educational policy changes
____Other (please specify):

____
____

2. How will you disseminate the results of the trial? (please check all that apply):

____Press conference
____Press release
____We will distribute ETS' reports
____Other (please specify):

____
____

3. Are you interested in linking the State NAEP to existing assessments in your state?

____Yes ____No

Comments:____
____

4. Are you interested in linking the State NAEP to international assessments?

____Yes ____No

Comments:____
____

= =

PART D. Finally, we would like to ask you about your opinions about the future of the state-by-state NAEP program.

1. Will your state participate in the 1992 Trial State NAEP?

____Yes, definitely
____Yes, pending budgetary approval
____Probably will participate
____Probably won't participate
____Definitely won't participate

2. If your state did participate in the 1990 Assessment, but will not participate in 1992, which of the following reasons apply?

____We cannot afford the cost of the expanded 1992 assessment
____We were unhappy with the way the 1990 assessment was conducted
____Other (please specify):

____
____

3. Suppose that a state-by-state assessment were implemented at all three age/grade levels in several content areas, every other year. Without implying any commitment, what is your best guess about the amount your state might contribute on an ongoing basis?

____No more than half of what we spent (contributed) in 1990
____About as much as we spent (contributed) in 1990
____About twice what we spent (contributed) in 1990
____Three times what we spent (contributed) in 1990, or more

4. Are there changes in the administration of the Trial State Assessment that you would like to see implemented in 1992?

____Yes ____No

Please specify changes:____

5. If you answered yes to Question 4 above, have you suggested any of these changes to NAEP (NCES, ETS, Westat, NAGB) personnel?

____Yes ____No

Comments:____

6. Are there changes in the content of the Trial State Assessment that you would like to see implemented in 1992? (n.b., current plans are to test fourth and eighth grade math and fourth grade reading)

____Yes ____No

Please specify changes:____

7. If you answered yes to Question 6 above, have you suggested any of these changes to NAEP (NCES, ETS, Westat, NAGB) personnel?

____Yes ____No

Comments:____

8. Do you think Congress should pass legislation to implement a full blown (non-trial) state-by-state assessment in 1994? ____Yes ____No

9. If you answered "no" to the previous question (8), do you think that the 1994 assessment should be another trial assessment? ____Yes ____No

10. What grade levels and content areas would you like to see on the 1994 state-by-state assessment (whether it is a trial assessment or not)?

4th grade	____math	____reading	____other (please specify)____________
8th grade	____math	____reading	____other (please specify)____________
12th grade	____math	____reading	____other (please specify)____________

Do you have any other comments?______________________________

Are you interested in receiving a summary of the results? ____Yes ____No

Your Name: ____________________________

Position: ____________________________

State: ____________________________

Phone Number: ____________________________

THANK YOU AGAIN FOR YOUR HELP IN COMPLETING THIS SURVEY!

1990 TRIAL STATE NAEP EVALUATION
STATE TESTING DIRECTORS QUESTIONNAIRE
(for Non-Participating States)

JULY, 1990

The following survey is being conducted by the American Institutes for Research, acting as a subcontractor to the National Academy of Education, as part of its evaluation of the 1990 Trial State National Assessment of Educational Progress. Please complete the following questions, answering as accurately, completely, and honestly as possible. Be assured that all individual responses will be held in strict confidence; only aggregate data will be reported. Thank you very much for your cooperation!

1. Why didn't your state participate in the 1990 NAEP Trial State Assessment (TSA)? (please check all that apply)

____We thought the Trial would be too expensive
____The State Superintendent did not want to participate
____Did not receive enough information about TSA from NAEP
____State could not commit funds for the program at sign-up time
____We already participated in National NAEP, and therefore felt constraints against participating in the TSA
____Don't feel need for information provided by the TSA
____Would be too much extra burden on schools and staff
____Afraid TSA would interfere with regular state assessment program
____Regular state assessment program provides all achievement information we need
____Other (please specify):
____
____
____

2. Are there changes that could be made to the State NAEP that would increase the likelihood of your participation? (please check all that apply)

My state would participate if:

____NAEP paid for the cost of the local administrators
____NAEP paid for staff time for the paperwork involved in the administration
____NAEP provided school or district-level information
____NAEP could be linked in some way to our existing state assessment
____Other reason (please specify):
____
____
____

3. Will your state participate in the 1992 Trial State NAEP?

____Yes, definitely
____Yes, pending budgetary approval
____Probably will participate
____Probably won't participate
____Definitely won't participate

4a. Do you think Congress should pass legislation to implement a full blown (non-trial) state-by-state assessment in 1994?

____Yes ____No

4b. If you answered "no" to the previous question (8), do you think that the 1994 assessment should be another trial assessment?

____Yes ____No

5. What grade levels and content areas would you like to see on the 1994 state-by-state assessment (whether it is a trial assessment or not)?

4th grade	____math	____reading	____other (please specify)__________
8th grade	____math	____reading	____other (please specify)__________
12th grade	____math	____reading	____other (please specify)__________

Do you have any other comments? ____________________

Are you interested in receiving a summary of the results? ____Yes ____No

Your Name: ____________________

Position: ____________________

State: ____________________

Phone Number: ____________________

THANK YOU AGAIN FOR YOUR HELP IN COMPLETING THIS SURVEY!

Information on the Reported Costs

In the 1990 survey, your state report the following figures for the direct and indirect costs of the state and local component of the Trial State NAEP project:

	Direct Cost	Indirect Cost	Certainty of Estimate
State Level Costs:	______	______	______
Local Level Costs:	______	______	______

1. How were these figures determined? What are the individual components of the direct and indirect costs? What ingredients (cost components) are included in each cost estimate? (Provide separate explanation of each.)

   State
   Direct Costs: ______________________________

   Indirect Costs: ______________________________

   Local
   Direct Costs: ______________________________

   Indirect Costs: ______________________________

2. Do these figures reflect (Circle response):
   a. budget allocations
   b. actual expenditures?

   Explain: ______________________________

3. Were the funds allocated by the state sufficient to cover all costs? Yes No

4. Would you expect the resource ingredients incorporated into the worksheets above to add up to the cost figures reported in this survey?

Yes No

If not, why?

____

____

____

5. How might the costs of the State NAEP operations be reduced (state or local level)?

____

____

____

____

____

The Content and Curricular Validity of the 1990 NAEP Mathematics Items: A Retrospective Analysis

Edward A. Silver
Patricia Ann Kenney
Leslie Salmon-Cox
Learning Research and Development Center
University of Pittsburgh

October 1991

Author Notes

The cooperation and assistance of individuals at the Council of Chief State School Offices, the Educational Testing Service, the National Center for Education Statistics, and American Institutes for Research are acknowledged. Especially noteworthy was the generous assistance provided by Ramsay Selden (CCSSO), Chancey Jones and Steve Koffler (ETS), and Liz Hartka (AIR).

The authors also wish to acknowledge the valuable contributions of Carol Parke, who assisted in conducting and summarizing telephone interviews, and Patricia Siemasko, who labored to assemble the many pieces of this report.

Introduction

Since 1969, the National Assessment of Educational Progress (NAEP) has been monitoring the educational achievement of American students in a variety of subject areas and changes in the achievement levels over time. To date, there have been five national assessments in mathematics. Although the 1990 mathematics assessment was in some ways similar to the four that preceded it, this assessment was different in some important ways.

Like its predecessors, the 1990 NAEP mathematics assessment was organized around a set of mathematical process and content area objectives, referred to as a "framework," which guided the selection and development of the test items. The NAEP mathematics framework has evolved over time to reflect the "best thinking" of the educational community with respect to mathematics content and mathematical processes. Due to this natural evolutionary process, the mathematics objectives framework for the 1990 NAEP mathematics assessment reflected changes from the framework used in the 1986 and earlier assessments. Furthermore, the development of the 1990 mathematics framework was undertaken at about the same time as the publication of the National Council of Teachers of Mathematics (NCTM) *Curriculum and Evaluation Standards for School Mathematics*. The NCTM *Standards* reflect a vision of what it means to be mathematically literate in today's technological society, focusing heavily on problem solving, communication of mathematical ideas, critical reasoning, and connections to ideas and procedures both within mathematics and across other content areas. This vision was a sharp contrast to more traditional views of school mathematics.

A major difference between the 1990 NAEP mathematics assessment and prior NAEP assessments was the intention to report state-by-state results. In the prior assessments, results were reported only for the national sample and for various subgroups (e.g., sex, ethnicity, region). The Trial State Assessment, as it came to be called, tested mathematics at the grade 8 level in 1990, and was designed to provide state-level data that could be used for comparisons among participating states and territories and between individual states and national performance. Given the new framework and the intended state-level reporting of results, a consensual process in which teachers, curriculum specialists, mathematics educators, local school administrators, parents, and the general public were given an opportunity to provide feedback on the 1990 NAEP mathematics framework became even more critical than perhaps it had been for previous NAEP mathematics assessments.

This is the final report of a study designed to (a) examine the consensual process that produced the mathematics objectives framework used to select and develop items for the 1990 State Trial Assessments and (b) analyze the validity of the framework and the items themselves. The report is divided into three parts. The first two parts contain analyses of the consensual process and the mathematics framework and items, respectively. The third part contains a set of recommendations that emerge from the results of these two analyses.

Analysis of the Consensual Process

When the National Assessment of Educational Progress was authorized by Congress, the language of P.L. 98-511, Section 405(E) required that the goals and objectives for the assessment should be developed through a national consensual process, providing for the participation of teachers, curriculum specialists, school administrators, parents, and members of the general public. The consensual process for development of the 1990 NAEP mathematics assessment was coordinated by the Council of Chief State School Officers (CCSSO) and involved many individuals and groups, including an 18-member Steering Committee for the National Assessment Planning Project, an 11-member Mathematics Objectives Committee, mathematics specialists from 39 state departments of education, and a group of about 25 other mathematics education experts. Appendix A contains a listing of these individuals.

In this study the adequacy of the consensual process was examined with respect to the constituencies represented and the nature of the advice sought, the extent to which the advice and input affected the design of the mathematics objectives framework, and the extent to which the framework appeared to represent a true consensus of opinion among professionals in the field of mathematics education. Moreover, the state-by-state comparisons included in the 1990 NAEP mathematics assessment increase the need for a careful examination of the consensual process in at least two areas: the inclusion and consideration of national, state, and local school district perspectives on mathematical goals and objectives, and the opinions of state and local mathematics educators regarding the content of the test.

Data Sources

Archival records and other written materials (e.g., minutes of committee meetings, records of written comments sent to the CCSSO) were used to gather preliminary information about the development of the 1990 NAEP mathematics framework. Two important pieces of information emerged from the perusal of these materials: the timeline for the development of the trial state assessment model for NAEP, and a verification of the individuals and groups who were involved in the consensual process over its four-year history. Knowing the timeline facilitated an investigation of the structure of the consensual process, while the information from letters, meeting minutes, and the state-level surveys assisted in the identification of those individuals, groups, and states that seemed to contribute most to the process.

However, for the purposes of this study, a series of semi-structured interviews with key participants was the major source of information about the adequacy of the consensual process itself. Appendix B contains a complete listing of those who were included in the modest, yet representative, interview sample for this study. Selection of the interviewees was accomplished through two different methods. Because this was the first mathematics assessment aimed at state-level reporting and comparisons, one part of the interview selection process focused on state-level mathematics specialists–consultants who had participated in the consensual process. Through the various documents provided by the CCSSO, we identified 41 people (representing 39 states) who had provided feedback on the proposed 1990 NAEP mathematics framework, and then we sent each person a letter, a copy of which appears in Appendix C. The letter indicated the objectives of this study and the questions to be pursued; it stated that a sample of those who participated in the state-level consensual process would be chosen for telephone interviews; and it invited anyone who wished to be sure that their views were known to call us directly. As a result of this correspondence, we received two "self-nomination" requests from individuals who wished to participate and one response from an individual who preferred to respond in writing. Subsequently, semi-structured telephone interviews were conducted with a total of 10 state-level mathematics specialists–consultants (the 2 volunteers and 8 others selected at random). Appendix D contains a copy of the telephone interview protocol. The written response received from the one specialist–consultant was also included in the analysis.

The other part of the interview process used a sample of people who were involved at the CCSSO level, who served on one or more of the CCSSO committees, or who were designated as special consultants to the consensual process. This selection resulted in interviews with 18 individuals: 3 people who were involved in the process at the CCSSO level, 2 representatives from Educational Testing Service (ETS), 3 members of the Steering Committee, 4 members of the Mathematics Objectives Committee, and 6 mathematics educators who served as special consultants. Questions similar to those used in the telephone interviews were asked of these 18 individuals; however, the questions were revised to make them more appropriate to the person's role in the consensual process.

Findings Regarding the Consensus-Building Process

Inclusion. The CCSSO appears to have made a "good faith" effort to include the relevant constituencies in the process of developing a broad consensus. The major committees reflected membership from a broad array of organizations and consequently represented a large number of constituencies, including not only those specifically interested in mathematics education but those interested in education more generally. Although representation was broad, the CCSSO may have been negligent in failing to include the representation of minority organizations with an interest in the content of the NAEP mathematics assessment. It is somewhat surprising that major national organizations associated with mathematics education, such as the National Council of Teachers of Mathematics (NCTM), the American Association for the Advancement of Science (AAAS), or the Mathematical Sciences Education Board (MSEB) were not included officially in the process of building consensus. Despite the lack of inclusion of representatives of these organizations on major committees, it is clear that some representatives and publications of these organizations were consulted individually during the consensus-building process.

Looking across the states sampled in our interviews, the groups assembled to review the framework also reflected diverse areas of expertise and interest. However, no systematic representational pattern was evident in the groups assembled at the state level. For example, one state did not convene a group but handled the review via mail and telephone; a few other states used intact groups that had been formed for other purposes, such as reviewing the state curriculum guidelines; other states created ad hoc groups for the purpose of reviewing the NAEP mathematics framework. Given the brief time available for review, there was less planfulness than one might wish to have in an important review process. Teachers were often included in the state-level review groups, but one state review group excluded teachers because of difficulties presented by the time in which the review was conducted. Administrators, testing experts, and mathematics supervisors were included in some but not all of the state-level review groups; non-educators, such as industry representatives, were included in the review group for only one state. The impression drawn from the interviews is that a good-faith effort was made to include all relevant constituencies, but the actual inclusion of different constituencies was uneven.

Consensus on the mathematics framework. In January 1988, the CCSSO sent a survey instrument to the state mathematics specialists and directors, asking them to convene a review committee to comment upon the proposed 1990 NAEP mathematics framework. The instrument included a five-page questionnaire, a summary of the five content areas (arithmetic; measurement; geometry; data analysis, statistics, and probability; and algebra and functions), a set of sample items for grades 4, 8, and 12, and the proposed percentage distribution of exercises by grade and process and by grade and content. The state review committees were asked to respond to (a) the reasonableness of the content-by-process organization of the framework, (b) the inclusiveness of the content areas to be included in the assessment, (c) the appropriateness of the grade-level designations for each grade level, (d) the proposed distribution of the percentages of items in the process and content categories, and (e) the appropriateness of the type of forms used for the items (e.g., multiple-choice, open-ended). The committee members also were asked to provide comments and/or summary statements on the proposed framework report. The state review committees had approximately two weeks to convene and respond to this survey. A total of 39 state committees completed and returned the survey instrument to CCSSO.

This study of the consensual process included telephone interviews of a subset of those people who responded to the CCSSO survey. In general, the individuals interviewed felt that the mathematics objective framework for the 1990 NAEP assessment represented a reasonable compromise between current instructional practices and the recommendations of professional organizations. For example, the state-level mathematics specialists-consultants generally reported that their review committees did not express major discontent with the framework for the 1990 assessment. In most cases the 1990 NAEP mathematics framework was viewed as "a move in the right direction," and most of those interviewed expressed a hope that the framework and ensuing assessments will continue to move progressively. Particular pleasure was expressed about the de-emphasis on procedural skills in the 1990 framework. Cases in which there were disagreements about the framework appear to have involved quibbles about relative weightings of content areas or abilities or the grade-level placement of items rather than major disagreements about purpose or focus.

Given the format of the questions on the CCSSO survey and the tight timelines involved, it is not surprising that the people interviewed for this study of the consensual process reported little disagreement. Perusal of the survey instrument reveals that the state committees were asked to review a framework that was already generated; they were not specifically asked to create one. They were asked to react to a given set of mathematical content areas; they were not invited to propose new ones. In general, then, the state committees were not invited to be critical at the structural level; they were merely being reactive to a structure that was predetermined.

Two of the people who participated in our telephone interviews reported a much more negative reaction to the framework. In one state the reaction to the framework was negative because it did not push far enough beyond current instructional practice; in the other state a mathematics specialist–consultant felt the framework pushed too far beyond current practice. In both cases, there was hope that future changes in the framework or in instructional practices would lead to a closer alignment between the state's expectations and the content of the NAEP mathematics assessment.

Consensus on NAEP design and intent. Among the state-level people interviewed, there appeared to be general interest in an alternative to the current NAEP framework and approach, such as an approach that could be issue-driven and a design that allowed for sufficient numbers of related items in order to answer questions such as, "What do our students know and understand about geometry?" or "How successful are students in engaging in proportional thinking in various contexts?" In other words, there was consensus that the 1990 framework was a move in the right direction, but 8 of the 10 persons interviewed in this study expressed interest in a more focussed assessment, even if it meant a less comprehensive test at any given point in time.

Limitations of consensus. At the state level, respondents indicated that the CCSSO review process for the mathematics objectives framework was useful and appropriate for intrastate consensus building. Because, however, there was no opportunity for one state committee to know how the other state committees responded, no interstate consensus was developed. The need for interstate consensus is heightened by the use of NAEP state-level data for the purposes of interstate comparisons. Unless interstate consensus is reached, it will be difficult to get agreement from all states that a fair measure is being applied commonly to all states. This issue becomes even more important if NAEP is to be used as a precursor to a national achievement test.

Of course, the attainment of an interstate consensus that is more than a "least common denominator" set of objectives will be a challenge. In the minutes of its August

1987 meeting, the NAEP Steering Committee indicated its intention to avoid a set of objectives that simply represented the intersection of those found in the participating states. Their argument for a consensual process that is "comprehensive and inclusive" appears to suggest a set of objectives that consists of the union of those found in the states rather than the intersection. Nevertheless, interstate communication and consensus-building are likely to be important as states strive to improve their mathematics programs, since they can use the interstate conversations, as well as the NAEP framework and the state-level assessment results, as tools for improvement of their state mathematics programs.

Insufficient time. A lack of time during the entire development process, from the summer of 1988 through the following summer, was noted by almost all of those interviewed as a major negative factor in the process. Since "silence was taken as assent" and timelines were extremely short, it is possible that some negative views were submerged by the process. Sufficient time for the development and consensus building process is likely to become even more important as the assessment is expanded in later years to include other grade levels and other subjects.

Analysis of the Validity of the Framework and Items

The mathematics objectives framework which guided the selection and construction of items for the 1990 NAEP assessment consisted of a 15-cell matrix (3 mathematical abilities by 5 content areas). The mathematical abilities and their associated weightings for the 1990 assessment were:

Conceptual understanding (CU)	40%
Procedural knowledge (PK)	30%
Problem solving (PS)	30%

The content areas and associated weightings for grade 8 were:

Numbers and operations (NO)	30%
Measurement (MT)	15%
Geometry (GY)	20%
Data analysis, statistics and probability (DA)	15%
Algebra and functions (AF)	20%

The framework appears in Figure 1.

Figure 1. Framework for the 1990 NAEP mathematics assessment

Mathematical Ability	**Content Area**				
	Numbers and Operations (NO)	Measurement (MT)	Geometry (GY)	Data Analysis, Statistics, and Probability (DA)	Algebra and Functions (AF)
Conceptual Understanding (CU)					
Procedural Knowledge (PK)					
Problem Solving (PS)					

In this study, the adequacy of the framework and the selected items was examined with respect to the identified ability and content categories, the intended weighted distribution of items by category, the actual distribution of items by category, and the sufficiency of numbers of items per cell for drawing inferences about student proficiency. Furthermore, attention was also given to the ease with which the framework can be understood by professionals in the field, the extent to which understandings or misunderstandings about mathematical goals or objectives are fostered by the framework, the range of important mathematical behaviors that students are required to perform in the assessment, the difficulty of the items, and the extent to which the framework and the items conform to contemporary thinking about mathematical goals for students. For example, the NCTM *Curriculum and Evaluation Standards for School Mathematics* places heavy emphasis on four major themes across all grade levels: (1) Mathematics as Problem Solving, (2) Mathematics as Reasoning, (3) Mathematics as Communication, and (4) Mathematical Connections. Of these, only the term *problem solving* is shared by both the NCTM *Standards* and the 1990 NAEP framework. Beyond assessing how well the NAEP items match the intended framework, it was important in this study to determine the extent to which the items adequately represent the themes identified in the NCTM *Standards*.

Data Sources

Official documents produced by NAEP and by the CCSSO describing the framework and the items provided a description of the framework and the content and ability categories. After the National Center for Education Statistics (NCES) transmitted copies of the 1990 NAEP mathematics items, we first determined the 7 "blocks" (i.e., sets of test items) that contained the subset of 137 items administered to grade 8 students for the state trial assessments. Next, classifications (i.e., mathematical ability and content) were obtained for each item. The NAEP item classifications were then matched against those proposed in the CCSSO-developed mathematics framework. Table 1 contains the intended vs. actual distribution of items by mathematical ability; Table 2 contains the distribution by content area; and Table 3 contains the ability-by-content distribution. As expected, with respect to both mathematical ability and content, the percentages of NAEP-classified items in each category and subcategory were approximately equal to that recommended in the framework proposed by the CCSSO Mathematics Objectives Committee.

Table 1. Intended (CCSSO) vs. actual (NAEP) distribution of grade 8 items by mathematical ability

(Percentages appear in parentheses.)

	CCSSO	NAEP
Conceptual Understanding (CU)	55 (40%)	57 (42%)
Procedural Knowledge (PK)	41 (30%)	43 (31%)
Problem Solving (PS)	41 (30%)	37 (27%)

Table 2. Intended (CCSSO) vs. actual (NAEP) distribution of grade 8 items by content area

(Percentages appear in parentheses.)

	CCSSO	NAEP
Numbers and Operations (NO)	41 (30%)	44 (32%)
Measurement (MT)	21 (15%)	21 (15%)
Geometry (GY)	27 (20%)	26 (19%)
Data Analysis, Statistics, and Probability (DA)	21 (15%)	19 (14%)
Algebra and Functions (AF)	27 (20%)	27 (20%)

Table 3. Intended (CCSSO) vs. actual (NAEP) distribution of grade 8 items by mathematical ability and content area

(Because of rounding, percentages may not add to 100 percent.)

	Content Area									
Mathematical Ability	Numbers and Operations (NO)		Measurement (MT)		Geometry (GY)		Data Analysis, Statistics, and Probability (DA)		Algebra and Functions (AF)	
	CCSSO	NAEP	CCSSO	NAEP	CCSSO	NAEP	CCSSO	NAEP	CCSSO	NAEP
Conceptual Understanding (CU)	17 (12%)	19 (14%)	8 (6%)	8 (6%)	11 (8%)	12 (9%)	8 (6%)	8 (6%)	11 (8%)	10 (7%)
Procedural Knowledge (PK)	13 (9%)	14 (10%)	6 (4%)	9 (7%)	8 (6%)	4 (3%)	6 (4%)	7 (5%)	8 (6%)	9 (7%)
Problem Solving (PS)	13 (9%)	11 (8%)	6 (4%)	4 (3%)	8 (6%)	10 (7%)	6 (4%)	4 (3%)	8 (6%)	8 (6%)

The framework and the items were then subjected to more rigorous scrutiny by project staff and by an Expert Panel (EP) of eight mathematics education professionals representing different constituencies and perspectives. Appendix E contains brief biographical sketches of the EP members. The EP met for two days in March 1991 to examine the adequacy of the NAEP framework and its relevance to current thinking in the field, and to examine the agreement between the framework and the items and between the intended and actual distribution of items by category in the framework. In addition, EP members also examined the items with respect to some alternate categories emerging from the NCTM *Standards* (e.g., communication, reasoning, connections) and offered comments and suggestions regarding the reporting of NAEP findings and the design of future assessments.

EP Analysis of the NAEP Framework and Items

Procedure. On the first day of the meeting, the EP discussed the 1990 NAEP mathematics framework. Two EP members, who had also participated as members of committees involved with the process of developing the 1990 NAEP mathematics assessment, provided clarification about the intended meaning of the terms used to describe the content and ability categories.

After extensive discussion of the categories and the meaning of terms, the EP members examined 15 sample items taken primarily from the NAEP booklet *Mathematics Objectives, 1990 Assessment* and matched these sample items with the framework categories. Following a discussion of the results of this practice activity, six EP members participated in the actual content validity matching. To avoid any appearance of a possible conflict of interest, the two EP members with prior involvement in the process of creating the 1990 NAEP framework or in the item development process did not participate in the validity study.

The content validity matching exercise involved the seven blocks used for the 1990 NAEP state trial assessment. The six EP members worked in pairs, with each pair classifying the items from three of the seven blocks. The items in the trend block (i.e., the set of 23 items that was administered in both 1986 and 1990, and then used to link performance across assessments) were classified by all EP pairs, in order to provide a measure of interrater agreement. Each pair was instructed to reach a consensus judgment regarding a classification for each item in their three blocks with respect to both mathematical ability and mathematical content. All pairs were able to complete the task within the two-hour period that was allotted. The procedures used in this study provide a more rigorous test of the "goodness of fit" of the items to the framework than the process used by the NAEP contractor, in which consultants simply agree or disagree with the classification supplied for a set of items.

Findings. Results from the trend block were used as a check on interrater reliability among the pairs of EP members. Of the 23 items, the pairs agreed completely on 18 items (78 percent). The five items that elicited disagreement involved one in content

(NO vs. AF) and four on mathematical ability (one CU vs. PS and three CU vs. PK). For the 18 items on which the Expert Panel pairs agreed with each other, 15 matched the actual NAEP classifications (83 percent). Of the three items on which the EP disagreed with NAEP, the disagreement was on items that NAEP classified as CU; the Expert Panel classified these items as PK. An example of one of these items appears in Figure 2. Although the NAEP mathematical ability classification was conceptual understanding, the EP members thought it involved the procedural knowledge associated with finding a common factor. Thus, given the rather high interrater agreement across the EP pairs (78 percent), there is reason to be confident that the ratings for items in the six non-trend blocks fairly accurately represent the judgment of the entire EP, even though only one pair made the classification.

Figure 2. Conceptual understanding (NAEP) vs. procedural knowledge (EP)

What number is a common factor of 10 and 15?

A) 5
B) 10
C) 15
D) 30
E) 150
F) I don't know.

Table 4 contains the results of the EP classifications of items from all seven blocks with respect to mathematical ability. The diagonal elements in this table show that the percent agreement between the NAEP and the EP was quite high for the PK-PK classification (88 percent) and PS-PS classification (76 percent). However, the agreement for CU-CU was only 51 percent, and the corresponding percent disagreement in the CU-PK (i.e., NAEP classification as CU; EP classification as PK) was 32 percent. These data suggest a repetition of the disagreement that emerged in the interrater reliability analysis of the trend block items. In both cases the EP members often classified items as PK in contrast to the NAEP classification of those items as CU. This tendency is evident in the actual number of items in these categories as well. The EP classified 33 items as CU; in contrast, NAEP classified 57 items as CU. The reverse pattern occurred in the

BEST COPY AVAILABLF

classification of PK items, with the EP classifying 62 items in this category and NAEP only 43.

Table 4. Percent agreement between NAEP and Expert Panel (EP) classification of grade 8 items: mathematical ability

(Diagonal [percent agreement] and off-diagonal [percent disagreement] entries based on NAEP row totals for CU, PK, and PS. Column totals for CU, PK, and PS represent the EP classification of the items. Because of rounding, percentages may not add to 100 percent.)

		Expert Panel (EP)			
		CU	PK	PS	No. of items: NAEP (% of total)
NAEP	CU	51%	32%	17%	57 (42%)
	PK	1%	88%	10%	43 (31%)
	PS	8%	16%	76%	37 (27%)
	No. of items: Expert Panel (% of total)	33 (24%)	62 (45%)	42 (31%)	137 (100%)

The data in Table 5, however, show a high level of agreement between the NAEP item classifications and those of the EP with respect to content. At the item level, the EP members disagreed with NAEP content classifications on only 14 of the 137 items used in the 1990 state trial assessment.

Table 5. Percent agreement between NAEP and Expert Panel (EP) classification of grade 8 items: content

(Diagonal [percent agreement] and off-diagonal [percent disagreement] entries based on NAEP row totals for NO, MT, GY, DA, and AF. Column totals for NO, MT, GY, DA, and AF represent the EP classification of the items. Dashed lines indicate empty cells. Because of rounding, percentages may not add to 100 percent.)

		Expert Panel (EP)					
		NO	MT	GY	DA	AF	No. of items: NAEP (% of total)
NAEP	NO	91%	–	–	2%	7%	44 (32%)
	MT	–	95%	–	–	5%	21 (15%)
	GY	4%	12%	85%	–	–	26 (19%)
	DA	5%	–	–	95%	–	19 (14%)
	AF	7%	7%	–	–	85%	27 (20%)
	No. of items: Expert Panel (% of total)	45 (33%)	25 (18%)	22 (16%)	19 (14%)	26 (19%)	137 (100%)

The high level of disagreement generated by the item classification on mathematical ability invites a discussion of causality. One probable cause of the disparity in the classification of CU and PK items is the presence of some ambiguity in the definitions of *conceptual understanding* and *procedural knowledge*. Although definitions of each were provided in both the CCSSO and NAEP contractor documents that described the 1990 NAEP mathematics framework, the terms themselves may be so interrelated that they cannot be thought of in isolation. As the sample item in Figure 2 illustrates, finding a common factor involves both conceptual understanding—the existence and nature of a common factor—and procedural knowledge—the process used to find such a factor.

Overall, on the classification of mathematical ability by content, the EP members agreed with the NAEP classifications on 87 items (64 percent). Table 6 displays both the NAEP and the EP classification of items within the mathematics framework itself. For example, in the CU-NO cell 19 items were given this classification by NAEP, and the EP members only judged 11 items as belonging to this category. In this case, the 11 EP-classified items were completely contained in the set of 19 NAEP-classified items. This pattern of total containment of the smaller number of items within the larger number appeared in two other cells (CU-MT and PK-MT). Disagreement on the classification of a total of only 20 items (out of 137) caused the lack of complete containment in the other 12 cells. For example, of the 4 items classified by NAEP as PS-MT, only 2 appeared in the set of 4 so classified by the EP.

Table 6. NAEP vs. Expert Panel (EP) distribution of grade 8 items by mathematical ability and content

(Because of rounding, percentages may not add to 100 percent.)

	Content Area									
Mathematical Ability	Numbers and Operations (NO)		Measurement (MT)		Geometry (GY)		Data Analysis, Statistics, and Probability (DA)		Algebra and Functions (AF)	
	NAEP	EP	NAEP	EP	NAEP	EP	NAEP	EP	NAEP	EP
Conceptual Understanding (CU)	19 (14%)	11 (8%)	8 (6%)	3 (2%)	12 (9%)	11 (8%)	8 (6%)	4 (3%)	10 (7%)	4 (3%)
Procedural Knowledge (PK)	14 (10%)	24 (18%)	9 (7%)	18 (13%)	4 (3%)	3 (2%)	7 (5%)	9 (7%)	9 (7%)	8 (6%)
Problem Solving (PS)	11 (8%)	9 (7%)	4 (3%)	4 (3%)	10 (7%)	8 (6%)	4 (3%)	6 (4%)	8 (6%)	15 (11%)

The data in Table 6 reveal another pattern of disagreements between the EP members and the NAEP classifications. The 1990 NAEP mathematics framework that guided item development was presented in matrix form in which individual items must fit into one mathematical ability category and one content area category. This requirement of mutual exclusivity may be more suitable for some ability-by-content areas than for others. For example, in Table 6 if one reads across the CU row, disagreements occurred most often in the content areas of NO (19 NAEP; 11 EP); MT (8 NAEP; 3 EP), DA (8 NAEP; 4 EP), and AF (10 NAEP, 4 EP). Likewise, for PK items the differences most often appeared in the content areas of NO (14 NAEP; 24 EP) and MT (9 NAEP; 18 EP). Reading down the content category NO reveals differences between mathematical ability classifications. In this instance, EP members did not classify as many items as CU-NO as were identified by NAEP (19 NAEP; 11 EP); however, they classified more items as PK-NO than did NAEP (14 NAEP; 24 EP). These results support the notions that mathematical ability and content are somehow interdependent and that items may not neatly fit into a rigid matrix structure.

For PS the primary source of disagreement occurred in the content area AF, with the EP including more items in this cell than NAEP (8 NAEP; 15 EP). Figure 3 shows an item that illustrates this disagreement. Classified by NAEP as CU-AF, this item was designated as by EP members as PS-AF. The rationale for the PS classification was that to the typical grade eight student, this problem would be novel and would thus evoke a problem-solving strategy rather than any conceptual understanding of the multiplicative identity.

Figure 3. Conceptual understanding (NAEP) vs. problem solving (EP)

Which of the following sentences is true when any number is substituted for x?

A) $x + 1 = x$
B) $x + 0 = 0$
C) $x \cdot 1 = 1$
D) $x \cdot 1 = x$
E) I don't know.

EP Judgment of "Match" Between NAEP Items and NCTM *Standards*

Procedure. The second part of the Expert Panel meeting involved an examination of the assessment with respect to unintended outcomes. In particular, EP members were asked to look at the 1990 NAEP items with respect to the four major themes emphasized in the NCTM *Standards*: Problem Solving, Reasoning, Communication, and Connections. As a prelude to this exercise, the four themes were discussed by all of the members. Appendix F contains the excerpts from the NCTM *Standards* for grades 5-8 which were made available to the EP members. The EP noted that in the NCTM *Standards* Problem Solving and Reasoning are two separate themes, while the NAEP framework considered reasoning as part of the problem-solving process. This distinction was noted as a potential difficulty for the classification exercise.

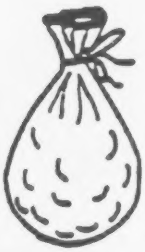
The eight Expert Panel members worked in pairs during the classification exercise; all EP members participated in this activity because no potential conflict of interest existed. Each EP pair classified items in two of the seven blocks, with two EP pairs each working with the trend block. Panel members were instructed to classify each item as Problem Solving, Reasoning, Communication, Connections, or None. In the case where more than one of the four themes were represented in an item, EP members were asked to rank order their classifications. EP members were asked to identify any items that they felt were particularly strong exemplars for any of the categories.

Findings. Before the overall analysis was done, the rating done by the pairs who had the 23 trend-block items was investigated for interrater reliability. Of the 23 items, the pairs agreed on 20 (87 percent) of them: 17 items (74 percent) did not address any of the four categories from the NCTM *Standards* and 3 items were classified as Reasoning (13 percent). The high degree of interrater agreement establishes confidence that the ratings done for other blocks by single rater pairs represent fairly accurately the view of the entire Expert Panel. Also, both rater pairs chose only one category for each of the items and did not need to use the rank ordering option.

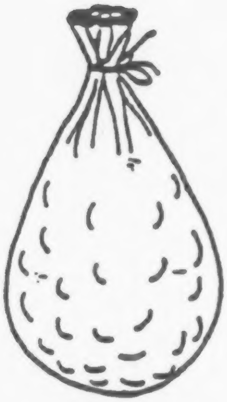
In the trend block, both rater pairs identified one item, shown in Figure 4, that was exemplary and classified it as Reasoning. In general, the trend block did not contain many items that met the criteria of the four NCTM *Standards* themes. This may not be surprising, since the trend block items came from the 1986 NAEP mathematics instrument, developed before the NCTM *Standards* existed. The vast majority of the items in the trend block were judged to match none of the four cross-cutting themes in the NCTM *Standards*; most of these items assessed students' computational proficiency.

Figure 4. Exemplary item: reasoning

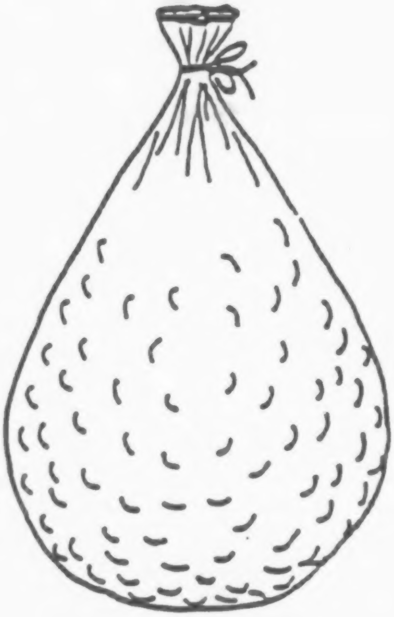
There is only one red marble in each of the bags shown below. Without looking, you are to pick a marble out of one of the bags. Which bag would give you the greatest chance of picking the red marble?



10 marbles



100 marbles



1000 marbles

A) Bag with 10 marbles

B) Bag with 100 marbles

C) Bag with 1000 marbles

D) It makes no difference.

E) I don't know.

The results of the EP matching of items from the six non-trend blocks with the NCTM *Standards* categories appears in Table 7. Because some rater pairs chose to use a rank-ordering system, the categories designated as the "first choice" were used to compile this table. This decision seemed to be reasonable in that only 25 of the 114 items were classified in two or more of the categories.

Table 7. Distribution of grade 8 items according to themes from the NCTM *Standards*

(Percentages appearing in parentheses are based on a total of 114 items from the 6 non-trend blocks.)

	Block						
NCTM *Standards* Theme	M4	M5	M6	M7	M8	M9	Total
Problem Solving	2	0	5	4	5	3	19 (17%)
Reasoning	11	10	7	3	1	4	36 (32%)
Communication	0	4	3	4	0	0	11 (9%)
Connections	0	1	0	0	0	0	1 (1%)
None	8	1	6	7	12	13	47 (41%)

More than half of the items were judged to be related in some nontrivial way to the major themes of the NCTM *Standards*, with the strongest relationship to the themes of Reasoning and Problem Solving. The number of items classified as being related to the NCTM *Standards* varied across the six item blocks, ranging from lows of 6 (out of 18) and 7 (out of 20) on two item blocks (i.e., Blocks M8 and M9, respectively) to a high of 15 (out of 16) on one block (i.e., Block M5). This latter item block consisted entirely of short-answer (i.e., non-multiple choice) items, which the EP members judged to measure best the themes from the NCTM *Standards*. Ironically, the two blocks with the lowest number of items judged as related to the NCTM *Standards* were those on which students were permitted to use calculators.

The issue of the calculator items deserves further elaboration. Calculators have been included in NAEP mathematics assessments since the 1977-78 administration. However, the 1990 NAEP assessment was the first time that the items themselves were actually classified in three ways: calculator-inactive (solution does not require the use of a calculator), calculator-neutral (use of calculator is left to the discretion of the student), and calculator-active (solution requires the use of a calculator). The number of items in each category in the two "calculator blocks" was: 16 calculator-inactive, 14 calculator-neutral, and 8 calculator-active. The EP members commented that the calculator-active items were not at all innovative; rather, they tended to involve only standard and uninteresting computations. In general, EP members judged calculator-neutral items to be more mathematically interesting that calculator-active items. An example of a calculator-active item from the grade 8 NAEP assessment appears in Figure 5.

Figure 5. Calculator-active item

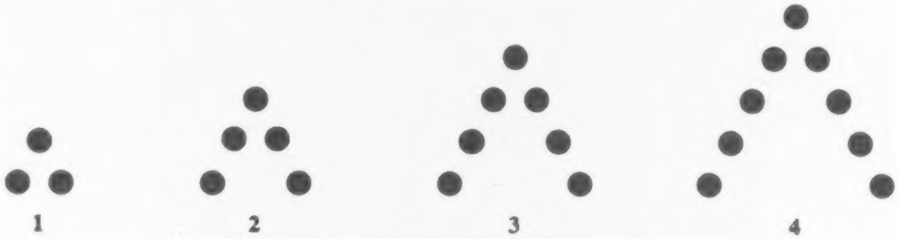
What is the product of 3.12 and 8^3?

Answer:

The EP members designated only 17 out of the 137 items (12 percent) as exemplars in particular categories. There was at least one exemplary item in each block; the block consisting entirely of short-answer items (M5) contained the largest number of exemplars (five). Two examples of exemplary items, one categorized as Problem Solving and Communication and the other as Reasoning, appear in Figure 6 and Figure 7, respectively. The item in Figure 7 was from one of the two calculator blocks, and was classified by NAEP as "calculator-neutral." The EP members commented that this item was far more interesting than any of the calculator-active items.

Figure 6. Exemplary item: problem solving and communication

The following questions refer to the following pattern of dot-figures.



If this pattern of dot-figures is continued, how many dots will be in the 100th figure? *[Problem Solving]*

A) 100
B) 101
C) 199
D) 200
E) 201

Explain how you found your answer. *[Communication]*

Figure 7. Exemplary item: reasoning

If 12 divides a whole number n without a remainder, list all whole numbers greater than 1 and less than 12 that must also divide n without a remainder.

Answers:

EP Discussion on the NAEP Framework and Items

Following each of the sessions described above, extensive discussion was held about the "matching" activities and more generally about the framework and the items. In particular, judgments of the EP members were sought regarding the overall adequacy of the framework, the intended and actual weighted distribution of items by category, the sufficiency of numbers of items per cell for drawing inferences about student proficiency, the range of important mathematical behaviors that students are required to perform in the assessment, the difficulty of the items, and the extent to which the framework and the items conform to contemporary thinking about mathematical goals for students. The four most salient points from the discussion are given below.

1. It was generally agreed that the framework was reasonable and represented an attempt to balance the vision represented in the NCTM *Standards* with the reality of current instructional practice in school mathematics. The major concern was that the framework not be viewed as final; subsequent refinements are needed to reflect, for example, an increasing convergence over time between the reality of math classroom practice and the vision of the NCTM *Standards*.

2. Concern was expressed about the adequacy of the number of items on the test. The number of items has decreased substantially since the original NAEP mathematics assessment, and the EP members felt that this severely limited the evidence base from which statements could be made about mathematical proficiencies in specific areas. A cautionary note was added, however, that simply adding more items will not automatically improve the test. Attention must also be given to issues of design and item format.

3. Despite the view that the framework was reasonable for the 1990 assessment, there was considerable concern about the limitations imposed by a "matrix view" of mathematics assessment. This limitation was referred to previously as one of the probable causes of the NAEP–Expert Panel disagreements concerning "conceptual understanding" and "procedural knowledge." Concern was expressed, for example, that use of a matrix fosters the notion that mathematics should be compartmentalized into discrete topics rather than thinking of mathematics as an integrated whole. This attention to a "topical laundry list" (as one EP member called it) focuses attention on narrow slices involving

disconnected, individual items instead of on mathematics and the test as whole entities. Moreover, EP members felt that the absence of tasks that matched the NCTM criterion for mathematical connections was due to the limitations imposed by "matrix thinking." That is, in constructing the test, one tends to focus on filling a cell with an item that fits in that cell and not elsewhere—each item has one and only one place in the matrix. This design constraint is in direct conflict with contemporary views of mathematics as a rich set of interconnected ideas.

4. Concern was expressed about the excessive use of multiple-choice items in the assessment. There were about three times as many multiple-choice as open-response items in the assessment; and many of the open-response items were not truly open-ended. The Expert Panel noted that the best fit between the NCTM *Standards* and NAEP items had been found for the open-response items. Moreover, the EP expressed dismay about the limited nature of the items on the assessment that allowed students to use calculators.

Recommendations

On the basis on our analysis of information gathered in this study and our own independent examination of the 1990 NAEP mathematics framework and test items, we offer the following set of recommendations.

1. The number of items on future NAEP mathematics assessments needs to be increased, and the items need to be carefully scrutinized for their validity as indicators of important mathematical knowledge and performance.

NAEP is designed to provide comprehensive coverage of topics at each grade level. Over the history of the assessment, the number of mathematics items has decreased substantially, due to a combination of economic factors (e.g., the high cost of new item production) and modifications in test design. There is reason to be concerned about the adequacy of the number of items on the test to provide an adequate picture of mathematical performance on the broad array of topics at a given grade level. In many cases, inferences will be made about the understanding or competence of the nation's students on some mathematics topics on the basis of performance on a very small number of items. Either the number of items needs to be increased, or the test needs to be altered to be less comprehensive, with a corresponding adjustment in public expectations regarding the comprehensiveness of the test.

Information currently available about the 1992 NAEP mathematics assessment suggests that the number of items has been increased for the next assessment. Although this is likely to be interpreted as good news in the mathematics education community, it is important to note that simply adding more items will not automatically improve the test. It is essential that attention also be given to issues of test design, item format, and validity. The 1992 NAEP mathematics item pool will need to be carefully examined with respect to all of these concerns, especially the extent to which the new items reflect the vision of the NCTM *Standards*.

2. Item formats on future NAEP mathematics assessments need to be more varied, including the increased use of open-response tasks, and the validity of the tasks and the scoring criteria will need to be carefully examined.

The format used for most of the 1990 assessment items was multiple choice, and it is clear that more varied task formats are needed in future assessments. The Expert Panel rated the few open-response items as being much more representative of current thinking about important objectives of mathematics education. Unfortunately, few innovative items were included in the state-by-state test. One type of innovative item format included in the 1990 assessment was the calculator-active item, in which students are allowed to use calculators that are provided to them. Unfortunately, the calculator items included in the 1990 assessment were quite unimaginative; they involved mostly straightforward calculations to be done with a calculator, not challenging tasks for which the calculator could be used as a tool for understanding or analysis.

Attention needs to be given to developing and administering a broader range of tasks that measure a wider variety of mathematical proficiencies and performances. Increasing the number of open-response tasks will, however, also increase the importance of attention to valid scoring of students' responses. The extent to which open-response mathematics tasks measure students' knowledge and ability with respect to language and communication rather than mathematics will become a more salient issue.

3. In designing future NAEP mathematics assessments, consideration should be given to alternatives to the content-by-process matrix structure to guide item writing and test construction.

Despite the view that the 1990 mathematics framework was reasonable, there was concern expressed by the Expert Panel and by some of those interviewed regarding the consensual process about the limitations imposed by a "matrix view" of mathematics assessment. Concern was expressed, for example, that use of a matrix to guide the development of test items fosters the notion that mathematics should be compartmentalized into discrete topics rather than thought of as an integrated whole. Certainly, the EP members felt that the absence of tasks matching NCTM's criterion for mathematical connections was due to the limitations imposed by a development process in which items are designed primarily to fill cells in a matrix rather than to measure good mathematical thinking in interesting ways. This approach to item construction not only stands in conflict with contemporary views of mathematics as a rich set of interconnected ideas, but it also can lead to confusions about item classification, as was discussed above in the analysis of the EP matching of items to the NAEP framework.

4. The design of the NAEP mathematics assessment should be examined to find ways to increase both its focus and its depth.

Current approaches to constructing the NAEP mathematics assessment are limited because the requirement of inter-item independence excludes "families" of related items. Clusters of related item—such as those contained in the early NAEP mathematics assessment—could, for example, yield deeper information about students' understanding of a mathematical "big idea" (e.g., volume or proportional thinking) or the understanding of integrated ideas (e.g., connections among algebraic, geometric, and arithmetic representations of multiplication). Consensus existed within the Expert Panel and also among the state representatives we surveyed that the value of the information obtained from NAEP would be increased if the design could be modified to allow increased focus and depth. Some clusters of related items were designed for earlier versions of NAEP, but because of the public release of some tasks, the limited time for new task development, and the constraints imposed by the current BIB design, these clusters are no longer part of the assessment.

One approach to increasing focus and depth would involve supplementing the current broad sampling of the entire mathematics curriculum with a more intensive examination of selected topics. Selected "focus topics" could receive greater attention in a given assessment. For the focus topics, a variety of task formats could be used to assess students' proficiencies and knowledge, and detailed accounts could be written about what the nation's students "know and are able to do." Moreover, the focus topics could vary across assessments and be repeated in some cyclical fashion, supplementing the other information available from NAEP at given points in time and providing an additional measure of changes in student achievement over time. It is likely that this approach would require both more money and more testing time. If neither was available, it might be wise to consider an approach that concentrated on focus topics in each assessment, varying them systematically across assessment occasions (e.g., geometry and proportional thinking in 1996 and 2006, rational numbers and deductive/inductive reasoning in 1998 and 2008), and abandoning the current approach that attempts to assess all areas of mathematics and succeeds in measuring only a thin veneer of mathematics each time.

5. The 1990 NAEP mathematics objectives framework was adequate for use in that assessment, but it must be viewed as an evolutionary entity and allowed to evolve to reflect changing conceptions of mathematics education in the nation.

Almost everyone interviewed or consulted during the conduct of this investigation agreed on the adequacy of the 1990 NAEP mathematics objectives framework. The designers of the 1990 framework apparently succeeded in developing a document that captured the spirit of change in mathematics education at the time it was written, yet preserved enough traditional elements to represent a reasonable compromise between the visions of educational reformers and the realities of educational practice.

There was also virtual unanimity in the view that the framework must be viewed as evolutionary, not fixed for long periods of time. The framework must be allowed to evolve to reflect changing conceptions of mathematics education in the nation, especially during these times of field-based reform. As the framework is modified, a consensual process needs to be maintained so that the framework can remain consistent with the views of education leaders at the state and national levels.

Despite general agreement on the need for an evolutionary perspective on the mathematics framework, opinions vary on the length of time that should be in the life cycle of a given framework. The NAEP mathematics framework was changed between the 1982 and 1986 assessments and then again between the 1986 and 1990 assessments. Now that the assessment will be given every two years instead of every four years, it may be impractical and unwise to alter the framework with each assessment. Nevertheless, a regular expectation of changes to the framework and a plan for making those changes are both needed. It is important, however, that adequate time be built into the revision cycle so that consensus can be maintained.

6. The amount of time for development, review, and critique in all aspects of the NAEP mathematics assessment must be increased.

One theme that emerged from almost every discussion about the 1990 NAEP mathematics assessment was the lack of adequate time available to complete the complex tasks that were required. For example, the timelines for framework development and review were reported to be too short. Time was also inadequate to allow for interstate consensus on the framework. Moreover, the timelines for item development, review, and field testing also appear to have been too short. A project done on such a tight schedule is bound not only to be conservative in its approach but also to depend extensively on past practices and existing structures.

Based on the results of this study, it is clear that planning must begin well in advance of future assessments so that adequate time can be allowed for the development of exemplary tasks to measure students' "mathematical power" and to operationalize the assessment of "mathematical connections." Especially in times of great reform efforts, sufficient time is needed to step back from development processes in order to reflect on both the activities and the products.

The recent decision to authorize administration of NAEP mathematics assessments every two years instead of every four years, due to pressures for more frequent national monitoring of student achievement, increases the likelihood that sufficient time will not be available for design, review, and critique in all aspects of the assessment. The increased tension between the need for increased time identified in this study and the increased frequency of administration of the assessment deserves careful monitoring in the future.

7. In order to reduce the likelihood of confusion and misinterpretation, the language used to frame the assessment and report the results needs to be clear and the terms used to describe mathematical content and process categories need to be clearly defined.

From the validity study it became clear that even experts in the field of mathematics education who had been carefully informed about NAEP's definitions of the terms *conceptual understanding* and *procedural knowledge* had a great deal of difficulty reaching agreement about the operationalization of those definitions in terms of item classification. One would expect even greater confusion when these terms are used by broader segments of the educational community and the general public. The frequent disagreements could have been the result of miscommunications and misunderstandings, or they could have resulted from fundamental differences of opinion about the mathematical proficiencies being tested. The former explanation certainly points to a need for increased clarity in defining terms. The latter case indicates a deeper, more substantive problem that goes beyond the need for clear definitions of terms, and it needs to be addressed both by those who report and interpret the results of the assessment and by those who design future assessments.

More attention to issues of language and communication during the consensual process might have avoided some of the confusions detected in this investigation. This need for attention should be noted for future assessments, and more time should be devoted to exploring the matters of language and communication during the drafting of future frameworks.

8. In future NAEP mathematics assessments a greater convergence is needed between the ideas and language used to frame the design of the test and those used in reporting the results of student performance.

A major issue to be faced in the future development of the NAEP mathematics assessment is the better alignment of the approaches used for test development with those used in reporting results. As has been the case for all NAEP mathematics assessments, the 1990 NAEP mathematics framework identified content and ability categories that were used to create a matrix; the matrix was then used to write the items for the test. Although the construction of the test was guided by the objectives framework, the framework (especially the mathematical ability categories) appeared to have little importance in framing the reporting of results.

For the 1990 NAEP mathematics assessment, results were reported as proficiency scales, which were derived post-hoc on the basis of student performance on the test. Proficiency scales, which were also used in reporting the results of the 1986 NAEP mathematics assessment, are quite useful in reporting student performance, but they are not closely related to the framework that was used to develop the items and construct the test. It is clear from discussions with Expert Panel members, and from other interviews conducted in this study, that substantial concern exists in the mathematics education community about the wisdom of reporting NAEP mathematics results by using scale scores. Since much of the reform movement in mathematics education is based on the assertion that mathematics is not unidimensional, the use of a single scale score to report mathematics proficiency appears to be incompatible with that view. On the other hand, the use of separate scale scores for each content strand is troublesome because it suggests a lack of connection across mathematical topics in different strands. The reporting of results using scale scores is also somewhat problematic because it tends to direct more attention to issues of content than to issues of abilities or processes. Higher scale scores result almost exclusively from the more difficult items and those dealing with more advanced or less commonly taught content. Moreover, the post-hoc descriptions of scale score levels lack a conceptual coherence that could help guide new task development or the improvement of instructional practice.

Beyond the specific concerns regarding scale scores is a more general concern about matching the ideas that guide the construction of the test more closely to those that guide the reporting of results. The introduction of post-hoc assignments of grade levels to the scale score levels (200 as grade 3, 300 as grade 7, and 350 as grade 12), as was done in reporting the 1990 results, is troublesome. The original item writers did not have in mind the intention of writing an item that measured grade 7 work and that would scale at 300 when they wrote the items that eventually formed the performance basis for that scale score. If item writers knew how the results were going to be reported and wrote items that were specifically designed for that purpose, then the items might have been different in character or in content. The same comment is applicable to the introduction by the National Assessment Governing Board of achievement levels (e.g., basic, advanced) in reporting student performance on the assessment. Without greater convergence between the ideas and language used to frame the design of the test and those used in reporting results, there is likely to be confusion and misunderstanding about appropriate conclusions that can be drawn from the assessment.

9. A coherent, sustained, reflective process of monitoring and shaping future NAEP mathematics assessments needs to be put into place immediately, and the process must be deeply informed by those most knowledgeable about school mathematics and education.

Given the importance of NAEP as an information source about students' mathematical knowledge and proficiency, it is important that knowledgeable persons from the field of mathematics education be included at all steps of the process. Although a strong effort was made to involve knowledgeable mathematics educators during the design of the framework for the 1990 NAEP, and though some effort was made to do this during item development and review, the lack of time for reflection and careful consideration of issues was a limiting factor to the value of such participation. Moreover, although a large number of different groups was involved in the process—on the CCSSO and ETS committees and as consultants—there was no single monitoring group with deep knowledge of the subject matter of the test to oversee the entire process.

In order to accomplish all that is needed for future NAEP mathematics assessments, a group of experts in mathematics education and assessment should be convened as soon as possible. They should be empowered to consider alternatives for NAEP design, content, task format, and scoring criteria so as to increase the value of the information obtained from this important national endeavor. It is also necessary that this group be charged with oversight of all the processes involved in the NAEP mathematics assessments and that the group be kept intact for long-term monitoring. Ad hoc groups formed for special purposes will also be needed, as they have been in all previous NAEP mathematics assessments, but it is vital that continuous oversight be provided by some group of knowledgeable content experts. Such a group should be able to give coherent, sustained, reflective attention to monitoring and actively shaping the NAEP mathematics assessment over the next decade, which will almost certainly be marked by major changes in mathematics education and assessment in this nation.

APPENDIX A

CCSSO COMMITTEE MEMBERS AND CONSULTANTS

Steering Committee

American Association of School Administrators

James E. Morrell - Superintendent of Public Schools, Muhlenberg, Pennsylvania

American Federation of Teachers

Antonia Cortese - First Vice President, State United Teachers, New York

Association of State Assessment Programs

Thomas Fisher - State Assessment Administrator, Florida

Association for Supervision and Curriculum Development

Alice Houston - Assistant Superintendent, Seattle Public Schools, Seattle, Washington.

Council for American Private Education and National Association of Independent Schools

Glenn Bracht - Director, American Lutheran Church, Minnesota

Council of Chief State School Officers

Richard A. Boyd - Superintendent of State Department of Education, Jackson, Mississippi

Council of the Great City Schools

Lillian Barna - Superintendent, Albuquerque, New Mexico

Directors of Research and Evaluation

Glenn Ligon - Director, Department Management/Information Service, Austin Independent School District, Texas

National Association of Elementary School Principals

C. June Knight - Principal, Hobart Middle School, Oklahoma

National Association of Secondary School Principals

Stephen Lee - Principal, Southwood High School, Indiana

National Association of State Boards of Education

Barbara Roberts Mason - President, State Board of Education, Michigan

National Association of Test Directors

Paul LeMahieu - Director of Research, Testing and Evaluation, Pittsburgh Public Schools, Pittsburgh, Pennsylvania

National Community on Catholic Education Association

Mary Brian Costello - Superintendent, Archdiocese of Chicago, Chicago, Illinois

National Council of State Legislators

Wilhelmina Delco - Congresswoman, Texas House of Representatives, Texas

National Education Association

Robert Astrup - President, Minnesota Education Association, Minnesota

National School Board Association

William M. Soult - Director, St. Urain Valley Board of Education, Colorado

National Governors' Association

Nancy DiLaura - Education Assistant, State House, Indiana

Mathematics Objectives Committee

Burks, Joan - Teacher, Damascus High School, Damascus, Maryland

Curtis, Phillip - Professor of Mathematics, University of California at Los Angeles, Los Angeles, California

Denham, Walter - Director of Mathematics Education, California Department of Education, Sacramento, California

Fisher, Thomas - Administrator of Assessment, Testing and Evaluation, Florida Department of Education, Tallahassee, Florida

Kahn, Ann - Past President, The National PTA, Fairfax, Virginia

Lindquist, Mary M. - Professor of Mathematics, Columbus College, Columbus, Georgia

Purser, Susan - Principal, Whitten Junior High School, Jackson, Mississippi

Strong, Dorothy - Director of Mathematics, Chicago Public Schools, Chicago, Illinois

Tucker, Thomas W. - Professor of Mathematics, Colgate University, Hamilton, New York

Watson, Charles - Mathematics Specialist, Arkansas Department of Education, Little Rock, Arkansas

Wells, Jr., R. O. - Professor of Mathematics, Rice University, Houston, Texas

State Education Agency Consultants

Auman, Anne - Mathematics Specialist, Kansas

Babb, James - Mathematics Professor, North Dakota

Badger, Elizabeth - Director Assessment Program, Massachusetts

Barlow, John - Curriculum Coordinator, Mississippi

Bassett, Judy - Curriculum Coordinator, South Dakota

Bennett, Milton - Mathematics Coordinator, West Virginia

Chambers, Donald L. - Mathematics Supervisor, Wisconsin

Colwill, Connie - Curriculum Coordinator, South Dakota

Comstock, Margaret - Mathematics Consultant, Ohio

Cowan, Peggy - Mathematics Specialist, Alaska

Danaher, June - Chief/Sciences and Mathematics, Maryland

Denham, Walter - Director of Mathematics Education, California
Dolan, Dan - Mathematics-Computer Education Specialist, Montana
Earl, William M. - Mathematics Specialist, Utah
Edwards, Jr., Edgar L. - Associate Director of Mathematics, Virginia
Erickson, John - Mathematics Specialist, Minnesota
Farley, Tom - Mathematics Consultant, Idaho
Fenton, Claire - Mathematics Consultant, New Mexico
Fineran, Don - Mathematics Specialist, Oregon
Gay, Susan - Mathematics Specialist, Oklahoma
Hanna, Karen - Mathematics Specialist, Tennessee
Henry, Renee - Mathematics Specialist, Florida
Kenney, Bob - Mathematics Consultant, Vermont
Leinwand, Steven - Mathematics Consultant, Connecticut
Long, Vera - Mathematics Consultant, Missouri
Martin, Phyllis - Mathematics Consultant, Georgia
Meek, Cleo - Mathematics Associate Director, North Carolina
Meeks, Wendell - Education Consultant, Illinois
Mitchell, Jacqueline P. - Mathematics Consultant, Maine
Nishimura, Kathleen - Mathematics Specialist, Hawaii
Paul, Fredric - Chief of Bureau Mathematics, New York
Peeler, Lane E. - Mathematics Specialist, South Carolina
Pledger, Linda - Mathematics Specialist, Alabama
Pollard-Cole, Mattye - Mathematics Specialist, Colorado
Prevost, Fernand J. - Mathematics Consultant, New Hampshire
Reeves, Andy - Mathematics Specialist, Florida
Vice, Sheila - Mathematics Consultant, Kentucky
Watson, Charles - Specialist of Mathematics, Arkansas
Wiley, Larry - Mathematics Curriculum Specialist, New Jersey
Wilson, Ruth - Mathematics Specialist, Nevada
Wilson-Hegg, Martha - Mathematics Consultant, Indiana

Other Consultants

Brown, Catherine - Virginia Polytechnic Institute and University
Campbell, Patricia - University of Maryland
Crosswhite, Joseph - Northern Arizona University
Dossey, John - Illinois State University
Gates, James - National Council Teachers of Mathematics
Hill, Shirley - University of Missouri
Hirsch, Christian - Western Michigan University
Kilpatrick, Jeremy - University of Georgia
Kouba, Vicki - University of Albany
Lappan, Glenda - Michigan State University
Lichtenberg, Betty - University of South Florida
Perry, Marcia - Georgia State University
Porter, Andrew - Michigan State University
Ralston, Anthony - University of New York
Romberg, Thomas - University of Wisconsin
Rosen, Linda - Mathematics Sciences Education Board
Schwartz, Judah L. - Harvard Graduate School of Education
Steen, Lynn Arthur - St. Olaf College
Stiff, Lee - North Carolina State University
Swafford, Jane - Illinois State University
Trafton, Paul - National College of Education
Travers, Kenneth - University of Illinois
Usiskin, Zalman - University of Chicago

APPENDIX B

PEOPLE INTERVIEWED FOR THE STUDY EXAMINING THE CONSENSUAL PROCESS

CCSSO Representatives

Wilmer Cody - Director, CCSSO State Assessment Development, 1988-1989
Edward Roeber - Member, CCSSO staff, 1988
Ramsey Selden - Director, CCSSO State Education Assessment Center

Steering Committee Members

Antonia Cortese
Paul LeMahieu
Glenn Ligon

Objectives Committee Members

Walter Denham
Thomas Fisher
Mary Lindquist
Charles Watson

Other Consultants

Catherine Brown
F. Joe Crosswhite
Jeremy Kilpatrick
John Dossey
Thomas Romberg
Jane Swafford

State-Level Mathematics Specialists/Consultants

Elizabeth Badger, Massachusetts
Donald Chambers, Wisconsin
Walter Denham, California
John Erichson, Minnesota
Steven Leinwand, Connecticut
Phyliss Martin, Georgia
Cleo Meek, North Carolina
Lane Peeler, South Carolina
Mattye Pollard-Cole, Colorado
Charles Reeves, Florida
Charles Watson, Arkansas

NAEP-ETS Representatives

James Braswell; Senior Examiner, Educational Testing Service, College Board Programs

Chancey Jones; Principal Measurement Specialist, Educational Testing Service, College Board Programs

APPENDIX C

LETTER TO STATE-LEVEL MATHEMATICS SPECIALISTS/CONSULTANTS

«DATA State Ed. Agency Cons.Labels»

Date

«Title»«FirstName» «LastName»
«Address1»
«IF Address2»«Address2»
«ENDIF»«IF Address3»«Address3»
«ENDIF»«IF Address4»«Address4»
«ENDIF»«City», «ST» «Zip»

Dear «Title» «LastName»:

I am presently directing a study of the content and curricular validity of the 1990 NAEP state-level mathematics assessment. This study is part of a program of studies being carried out for the National Academy of Education. In my study, the objective is:

> *to assess whether (a) the 1990 NAEP mathematics items reflect advanced thinking and best practices in mathematics assessment and (b) the consensus process for developing the objectives was adequate. In particular, we want to know whether the 1990 mathematics items reflect the range of content and processes, and the pattern of emphasis represented by recently developed NCTM and MSEB frameworks for mathematics curricula.*

You--and people in positions similar to yours in other states--were integrally involved in the development process we are studying. Our records indicate that in January 1988 you called together a committee of people to critique the draft of the Framework for the mathematics assessment. We will be interviewing a sample of people from this group, because of the importance of that state-level review. We will want to know:

- What was your evaluation of the Framework when you saw it in 1988? How well did it reflect the philosophy/reality of mathematics education in your state at that time? How did you think it compared to previous NAEP exams as well as to ones you would like to see in the future?

- What kind of consensual process was used in your state? Who did you invite to critique the Framework and what did they think of it? Was there generally consensus, within the group, about the document?

- When you reported back to CCSSO (the organization which had sent out the Framework) how did you report disagreements? Did you hear from CCSSO again, after you submitted your report? In your opinion, were your views adequately considered by those responsible for creating the 1990 assessment?

- If you are now in a position to use the data generated by the 1990 state-level assessments in mathematics, do you think you will? If so, why and for what purposes? If not, why not?

We are conducting this study in a fairly brief time. Our plan therefore is to conduct telephone interviews with a sample of the state leaders of this process. However, if you have strong views on this subject, or if for whatever reason you want to be sure your views are part of our information base, I invite you to call and let us know your thinking. You may contact Dr. Leslie Salmon-Cox, who is assisting me with this study at 412-624-7477.

Sincerely,

Edward A. Silver

APPENDIX D

TELEPHONE INTERVIEW PROTOCOL

Telephone Interview Questions

As you may recall, I'm calling to get your views of a process that took place several years ago. That was the soliciting of your help in looking at the NAEP Framework for the State level mathematics assessment. Our records indicate that CCSSO contacted you first in January/February 1988. I wonder if you could share with me what you remember about what you were asked to do.

1. Who did you invite to critique the Framework?

Math educators____ State level___ Local district level___
Testing office people___ State level___ Locals____
Teachers____ Supervisors________ Administrators____School Board members____
Citizens____
Others (please note)

1.1 and what did they think of it?

2. Was there generally consensus, within the group, about the document?

3. How well did the Framework reflect the philosophy/reality of mathematics education in your state at that time?

4. What was your own evaluation of the Framework when you saw it in 1988--on a scale from 1 to 10, very negative to very positive, where would you rate your evaluation of it?

5. How did you think it compared to previous NAEP exams as well as to ones you would like to see in the future?

6. Did you hear from CCSSO again, after you submitted your report?

7. Did you, at that time, have any idea how your state's responses compared to those from other states? If yes, how? If no, would you have liked to have known, is that something that would have interested you?

8. If you are now in a position to use the data generated by the 1990 state-level assessments in mathematics, do you think you will? For what purposes? If not, why?

9. Would you care to comment on any relationships you see between the NAEP assessment, or Framework, and state level testing in your state?

Let me close with a speculative question. What you received to comment on back in 1988 was a Framework. That document led to the development of a matrix--encompassing content and process, which in turn, led to the development of specific items. There are other ways a NAEP assessment might be structured. For example, some people have talked of the desirability of a NAEP assessment that could really answer questions, one that was issue-driven. For example, there might be a clusters of items to answer the question, e.g., what do students know about geometry? or how good are students' estimation skills? or on what kinds of activities, from among a range, can students show proficiency with calculators? In other words, items would be developed to answer these questions rather than to fill out a matrix. Is this a kind of NAEP assessment that would interest you? others in your state? Do you think it would be possible to get consensus on such an assessment--in your state or nationwide?

APPENDIX E

BIOGRAPHICAL SKETCHES OF EXPERT PANEL (EP) MEMBERS

Dr. Gilberto Cuevas is a Professor of Mathematics Education at the University of Miami, Coral Gables. He has published extensively in the areas of mathematics education of language minority students, and has been recognized by the Mathematical Sciences Education Board for the work he has done with minority students in mathematics.

Ms. Connie Finseth is currently serving a Teacher-in-Residence for the QUASAR Project, based at the University of Pittsburgh's Learning Research and Development Center. Ms. Finseth has had over 20 years' teaching experience at the elementary and middle school levels.

Dr. Jeremy Kilpatrick, Professor of Mathematics Education at the University of Georgia, has served on many boards and committees, both at the national and international levels. His research interests include problem solving and abilities in mathematics as well as mathematics assessment. He has been involved with the NAEP mathematics assessment since its beginnings, and most recently has been a member of the Item Development Panel for the 1990 mathematics assessment.

Dr. Gerald Kulm is a Professor of Mathematics Education at Texas A&M University. Prior to assuming this position, he was Director of Science, Mathematics, and Technology Education programs at the American Association for the Advancement of Science. The author of many books, articles and reports on mathematics problem solving, teaching attitudes and learning, he is the editor of the recent AAAS publication, *Assessing Higher Order Thinking Skills in Mathematics*.

Dr. Frank K. Lester, Jr. is a Professor of Mathematics Education at Indiana University - Bloomington. His primary research interests lie in the areas of mathematical problem solving, especially with respect to problem-solving instruction at the elementary and middle-school levels. He is currently editor-elect of the *Journal for Research in Mathematics Education*.

Dr. Mary M. Lindquist, the Fuller E. Callaway Professor of Mathematics Education at Columbus College (GA), is the President-Elect of the National Council of Teachers of Mathematics (NCTM). She has had extensive involvement with the NAEP mathematics assessment over the years as a member of the NCTM interpretive team for the second, fourth, fifth, and sixth NAEP mathematics assessments, a co-author of The Nation's Report Card for the fourth NAEP mathematics assessment, and a member of the CCSSO committee that set the objectives for the 1990 NAEP mathematics assessment.

Dr. John Moyer is an Associate Professor in Marquette University's Department of Mathematics, Statistics, and Computer Science. In addition to teaching mathematics, computer science, and mathematics education courses, he is the director of the Metropolitan Milwaukee Mathematics Collaborative, the Middle School Teacher's Mathematics Project, and is co-director of the QUASAR partnership between Marquette University and Edison Middle School.

Ms. Mary Lynn Raith has extensive experience in middle school mathematics. For the past three years, she has served as a K-8 mathematics supervisor for the Pittsburgh Public Schools. Prior to assuming this position, she was a middle school mathematics teacher for fifteen years. Her supervisory responsibilities include curriculum development, design of alternative assessments, and the supervision of teachers and coordination of mathematics programs. Ms. Raith is also a doctoral degree candidate in Mathematics Education at the University of Pittsburgh.

APPENDIX F

THEMES FROM THE NCTM *STANDARDS* FOR GRADES 5-8

THEMES FROM THE NCTM STANDARDS (Grades 5-8)

STANDARD 1:
MATHEMATICS AS PROBLEM SOLVING

In grades 5–8, the mathematics curriculum should include numerous and varied experiences with problem solving as a method of inquiry and application so that students can—

- *use problem-solving approaches to investigate and understand mathematical content;*
- *formulate problems from situations within and outside mathematics;*
- *develop and apply a variety of strategies to solve problems, with emphasis on multistep and nonroutine problems;*
- *verify and interpret results with respect to the original problem situation;*
- *generalize solutions and strategies to new problem situations;*
- *acquire confidence in using mathematics meaningfully.*

STANDARD 2:
MATHEMATICS AS COMMUNICATION

In grades 5–8, the study of mathematics should include opportunities to communicate so that students can—

- *model situations using oral, written, concrete, pictorial, graphical, and algebraic methods;*
- *reflect on and clarify their own thinking about mathematical ideas and situations;*
- *develop common understandings of mathematical ideas, including the role of definitions;*
- *use the skills of reading, listening, and viewing to interpret and evaluate mathematical ideas;*
- *discuss mathematical ideas and make conjectures and convincing arguments;*
- *appreciate the value of mathematical notation and its role in the development of mathematical ideas.*

(From: National Council of Teachers of Mathematics. Curriculum and Evaluation Standards for School Mathematics, p. 75 ; 78.)

STANDARD 3: MATHEMATICS AS REASONING

In grades 5–8, reasoning shall permeate the mathematics curriculum so that students can—

- *recognize and apply deductive and inductive reasoning;*
- *understand and apply reasoning processes, with special attention to spatial reasoning and reasoning with proportions and graphs;*
- *make and evaluate mathematical conjectures and arguments;*
- *validate their own thinking;*
- *appreciate the pervasive use and power of reasoning as a part of mathematics.*

STANDARD 4: MATHEMATICAL CONNECTIONS

In grades 5–8, the mathematics curriculum should include the investigation of mathematical connections so that students can—

- *see mathematics as an integrated whole;*
- *explore problems and describe results using graphical, numerical, physical, algebraic, and verbal mathematical models or representations;*
- *use a mathematical idea to further their understanding of other mathematical ideas;*
- *apply mathematical thinking and modeling to solve problems that arise in other disciplines, such as art, music, psychology, science, and business;*
- *value the role of mathematics in our culture and society.*

(From: National Council of Teachers of Mathematics. Curriculum and Evaluation Standards for School Mathematics, p. 81 ; 84.)

The Relative Standing of States in the 1990 Trial State Assessment: The Influence of Choice of Content, Statistics, and Subpopulation Breakdowns

Robert L. Linn
Lorrie Shepard
Center for Research on Evaluation, Standards, and Student Testing
University of Colorado at Boulder

Elizabeth Hartka
American Institutes for Research

August 1991

Introduction

During most of the 1980s, state averages on either the American College Test (ACT) or the Scholastic Aptitude Test (SAT) were used as a primary educational "performance outcome" on the wall charts that were first introduced by former Secretary of Education Terrell H. Bell. Although the relative standing of states based on the average score of students taking one of these two college admissions tests received a good deal of attention in the press each year, the results were easily discredited as comparative indicators of educational quality for two major reasons and several lessor ones. First, the tests were designed for a different purpose and do not cover the content that would be desired for comparing the academic attainment of high-school seniors. Second, the tests are not taken by representative samples of students.

Dissatisfaction with the use of ACT and SAT results to compare states contributed to support for development of a better means of state-by-state comparisons through the use of the National Assessment of Educational Progress (NAEP) at the state level. The NAEP Trial State Assessment (TSA) program, created by an act of Congress, allowed for the voluntary participation of states in the 1990 mathematics assessment at grade 8 and the 1992 reading assessment at grade 4 and mathematics assessments at grades 4 and 8. The TSA is designed to obviate the two primary criticisms of the wall chart comparisons. The June 6, 1991 release of the grade 8 mathematics results from the 1990 TSA provided the first comparative results on a common achievement measure constructed on the basis of the NAEP mathematics framework for representative samples of the 40 TSA participants (37 states, 2 territories, and the District of Columbia).

Although the TSA effectively responded to the concern that previous wall-chart comparisons were distorted due to the non-representative nature of the groups of test takers and the measure is more suitable than college admissions tests for comparing student achievement, there are a number of other issues that can influence the relative standing of a state and the validity of inferences about the mathematical proficiency of its eighth grade students. Comparisons can be made in a wide variety of ways, and the choice of procedures used to make the comparisons can markedly alter the outcome. Hence, it is critical that the basis of comparison and sensitivity of relative outcomes to alternative approaches be evaluated.

The purpose of this report is to determine the degree to which the relative standing of a state is sensitive to the choice of scales, descriptive statistics, or groups of students within a state for which results are reported. To be more specific, the analyses reported were conducted in an effort to answer questions such as the following. How much difference does it make that the focus of attention is so narrowly confined to the state average in overall mathematics achievement? Would the use of subscores, alternative statistical summaries, different reporting metrics, or an emphasis on results for subpopulations lead to noticeably different relative standings for some states?

Factors Influencing Comparisons

Although state comparisons based on a common achievement measure for representative samples are new, the use of test scores to compare the performance of groups of test takers is not. Groups may be defined in terms of demographic characteristics of the test takers, the instructional program to which students are assigned, or in terms of administrative groups (e.g.,schools, districts, or states). Whatever the basis for forming the groups, average test scores typically provide the primary statistic for comparing group performance. Although such comparisons may appear straightforward and reasonable, they are, in fact, only one of many ways in which test results might be used to make group comparisons. Moreover, the relative standing of a group may vary substantially as the result of decisions that are made about test content and the way in which the group performance is summarized.

Comparative results in program evaluations have shown that the outcome can be quite sensitive to the choice of test content and the relative weight given to content categories (Walker and Schaffarzick 1974). The importance of choice of content and the weights assigned to components of composite scores was underscored by Cronbach (1971). **"If the original test or battery is a composite covering various types of content or various objectives, it implicitly weights those elements, either by the number of items allocated to each or by the way the score is calculated. Such a weighting cannot satisfy decision makers who hold values unlike those of the test developer"** (460, emphasis in original).

In addition to content influences, the choice of metric (e.g., percent correct, proficiency based on an item response theory scaling) can also change the relative standing of group averages (Spencer 1983). Moreover, even with the content and the metric fixed, the choice of descriptive statistics can radically alter group comparisons. An emphasis on the highest or lowest scoring students, for example, the 90th or 10th percentiles within each group, may order groups quite differently than the group averages.

Results for relevant subpopulations (e.g., racial-ethnic groups, parental socio-economic status) can also yield comparative results that seem to contradict overall averages. As demonstrated by Simpson's paradox (see, for example, Novick 1982), it is possible for the state A average to be higher than the state B average for each subpopulation (e.g., rural-urban, black-white-Hispanic) when the overall state average for state B is higher than the overall average for state A.

Simpson's paradox results from disparities from one population to the next in the proportion of people in the subpopulations and it is easy to construct examples of the paradox with fictitious data. As the actual results displayed in Table 1 for two of the states participating in the TSA illustrate, however, Simpson's paradox is more than a theoretical possibility. For all students, Texas had an average proficiency 2 points lower, albeit not significantly different, than Rhode Island. However, for each of the three racial-ethnic groups with sufficient sample size to permit reliable estimates, Texas students had higher average proficiency than their Rhode Island counterparts (by 8 points for white students, by 9 for black students, and by 17 for Hispanic students).

Cautions Regarding Simple Rank Order Comparisons

At the release of the 1990 National Assessment of Educational Progress (NAEP) mathematics results and the first-ever Trial State Assessment (TSA) results on June 6, 1991, reporters were urged not to use the results to produce a simple ranking of states. Emerson Elliott, for example, cautioned against simplistic ranking of states and stated that "improving education is the purpose, not the ranking of states." Gary Phillips explained the use of a figure that displays for each state the other states that have significantly higher averages, the states that have significantly lower averages, and the states where the average proficiency is not significantly different. He went on to say that the simple ranking of states without attention to the significance of differences was "incorrect, misleading, and irresponsible -- so don't do it."

Table 1. Illustration of Simpson's paradox using average overall mathematics proficiency for racial-ethnic groups for two states participating in the trial state assessment*

	Rhode Island		Texas		Difference
Racial/ Ethnic Group	Percent of Students	Average Proficiency	Percent of Students	Average Proficiency	Texas Minus R. I.
White	83 (0.8)	265 (0.8)	47 (2.1)	273 (1.1)	8
Black	5 (0.5)	225 (2.4)	13 (1.3)	234 (1.7)	9
Hispanic	8 (0.5)	228 (2.3)	36 (2.1)	245 (1.7)	17
Asian /Pacific Islander	2 (0.6)	*** (***)	2 (0.6)	*** (***)	***
American Indian	1 (0.2)	*** (***)	1 (0.2)	*** (***)	***
State Total	100	260 (0.5)	100	258 (1.3)	-2

* Racial-ethnic group percents and averages are from Table 12.1 (p.266) and total group averages are from Table 11.1 (p. 240) in Mullis, Dossey, Owen and Phillips 1991). The standard errors of the estimated percentages and average proficiencies are in parentheses.

*** Sample size insufficient to permit reliable estimate.

Similar cautions can be found in the full report of the TSA results. Standard errors and statistical significance of differences, based on the use of the Bonferroni procedure to account for the 780 possible pairwise comparisons among the 40 participants in the TSA, were stressed. The large number of pairwise comparisons that are not significantly different, is the basis for arguing that "it would be quite misleading to assign numerical rankings (1 to 40) to these results" (Mullis, Dossey, Owen, & Phillips 1991, 15).

Despite the pleas **not** to use rankings and efforts to suggest alternative approaches to reporting comparative results, the simple rank order based on average overall mathematics proficiency appeared in a number of news accounts of the TSA results. Relatively little attention was given to the variety of other statistics included in the report that could be used to compare the performance of states. Included among the many possibilities are results by five content categories, results based on alternative descriptive statistics (e.g., selected percentile points, and the percentage of students within each state scoring above the NAEP anchor points), results for subpopulations defined by gender, racial-ethnic group, type of community, parental education, results for students in the top and bottom one-third of the NAEP schools, and results based on teacher and student reports regarding mathematics instruction.

In the analyses reported here considerable attention is given to changes in rank order for states when the achievement data are summarized in different ways. The focus on rank order shifts should not be interpreted as an implicit endorsement of the use of a simple ranking of states to report the results of the TSA. On the contrary, the shifts in rank order that are demonstrated when different summaries of the results are used provide support for the need to look beyond any single rank ordering. More comprehensive use of the rich array of information about the distributions of performance of subpopulations within a state on different subscales is needed to describe eighth grade mathematics achievement within each state and to create the baseline for future assessments.

Method

The sensitivity analyses were designed to investigate the degree to which the rank ordering of states varies as a function of the definition of scales, the metric used for reporting, the choice of statistics, and the analysis of subpopulation results to characterize the performance of a state. Although there were 40 participants in the TSA, the analyses reported here were based on the 37 states. The overall performance of the District of Columbia and of the two participating territories, Guam and the Virgin Islands, was substantially lower than that of any of the 37 participating states. Excluding the District of Columbia, Guam, and the Virgin Islands from the analyses provides a better picture of the sensitivity of the rank ordering of states to choices of scales, metrics, statistical summaries, and the analysis of results for subpopulations. The analyses of data in each of these broad categories are briefly described.

Content and Process Scales

The TSA report (Mullis et al. 1991) includes results for an overall composite score, called "overall mathematics proficiency" and five content-specific subscores: numbers and operations; measurement; geometry; data analysis, statistics and probability; and algebra and functions. Thus, the main TSA report provides the basis for comparisons based on the composite and each of the five content scales. Other subscores formed by different combinations of items were also analyzed, however.

In addition to the five content categories, the NAEP mathematics framework categorized items according to a three-level process dimension: conceptual understanding, procedural knowledge, and problem solving. This process area breakdown provided the basis for a second set of scores used in the analyses. Categorizations based on a comparison of items to the National Council of Teachers of Mathematics (NCTM) Curriculum Standards completed by a study group headed by Edward Silver provided a third set of scores for the analyses (Silver 1991). Finally, comparisons were made based on scores computed separately for multiple-choice items and for open-ended items included in the assessment.

Metric

The TSA report used a proficiency scale based on item response theory. Five score estimates (called plausible values) for each of the five content areas, plus overall mathematics proficiency estimates were provided by Educational Testing Service (ETS). Plausible values are based on a combination of responses to the assessment items and a set of conditioning variables consisting of between 82 and 91 principal components of variables from teacher and student background questionnaires that were separately determined for each state (Mazzeo, 1991). The overall proficiency estimates, or composite scores, are weighted averages of the five content subscales, where the weights are proportional to the number of items in each subscale (Numbers and Operations, 45 items; Measurement 21 items; Geometry, 26 items; Data Analysis, Statistics, 19 items; and Algebra and Functions, 26 items). The plausible values were rescaled to the final proficiency scale metric using a linear transformation.

Percent correct scores and logits based on the percent correct scores were computed. The rank order of states using these scores was compared to the rank order using the ETS plausible value scoring technology.

Alternative Statistical Summaries

Several statistical summaries for the performance of each state were computed using the subscales and the composite. These include the state mean, selected percentile points of the state distribution of scores (10th, 25th, 50th, 75th and 90th percentiles), and the percentage of students scoring above the traditional NAEP anchor points of 150, 200, 250, 300, and 350. The rank orderings of states using each of these statistics were compared.

Subgroup Analyses

State results for several subpopulations were analyzed and compared to the results based on total state samples. Subpopulations were defined by racial-ethnic group, by the students attending the top third and bottom third of the schools within each state.

Results

Content and Process Scores

The use of state averages based on any of the five content subscores yields a rank order that is quite similar to that based on the state average on the overall mathematics proficiency composite. The rank-order correlations for the subscore and composite average scores are shown in Table 2. As can be seen, the rank-order correlation of each of the five subscore averages with the composite average was either .98 or .99. Even when one subscore is compared with another the rank-order correlations are all .95 or higher.

Distributions of differences in rank order for state means based on composite scores and state medians based on each of the five content subscores are shown in Figures 1 through 5. In only two cases (both involving the Geometry subscore) did the rank order of a state mean based on a content subscore differ by more than 5 ranks from the rank of the state mean on the composite score. For the majority of the states the rank order for a subscore average was within 1 rank of that for the composite. Based on high correlations shown in Table 1 and the relatively small differences in rank orders shown in Figures 1 through 5, it appears that state comparisons are not overly sensitive to the choice of content category or the relative weights that might be given to the five NAEP content categories in forming a composite score.

The creation of subscores based on the process categories in the NAEP framework (procedural knowledge, 41 items; conceptual understanding, 59 items; and problem solving, 37 items) produced state average scores that ranked states in nearly the same way as the average on the overall composite. As can be seen in Table 3, the average overall composite had a rank order correlation of .99 with each of the three average process subscores and the process subscores had rank order correlations of either .98 or .99 with each other. With only one exception the difference in rank orders based on the average overall composite and rank orders based on any of the three average process subscores was 3 ranks or less. The one exception differed by only 4 ranks for problem solving versus the overall composite.

Table 2. Intercorrelations among state ranks based on mean overall composite scores and each of the five content subscores

Variable	Overall Composite	Numbers and Oper.	Measure- ment	Geo- metry	Data Ana. & Stat.	Algebra & Functions
Composite	1.00	.99	.99	.98	.99	.99
Numbers & Operations	.99	1.00	.99	.95	.98	.98
Measurement	.99	.99	1.00	.97	.97	.98
Geometry	.98	.95	.97	1.00	.97	.97
Data Analysis, Statistics, & Probability	.99	.98	.98	.97	1.00	.99
Algebra & Functions	.99	.98	.98	.97	.99	1.00

N = 37 States.

Table 3. Intercorrelations among state ranks based on mean overall composite scores and each of the three NAEP process subscores

Variable	Overall Composite	Procedural Knowledge	Conceptual Understanding	Problem Solving
Overall Composite	1.00	.99	.99	.99
Procedural Knowledge	.99	1.00	.99	.98
Conceptual Understanding	.99	.99	1.00	.98
Problem Solving	.99	.98	.98	1.00

N = 37 States.

Figure 1. Distribution of difference in rank order of state mean based on overall mathematics composite and state mean based on the numbers and operations subscore

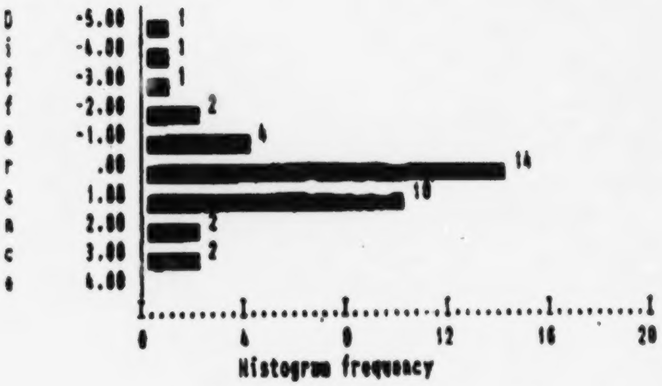


Figure 2. Distribution of difference in rank order of state mean based on overall mathematics composite and state mean based on the measurement subscore

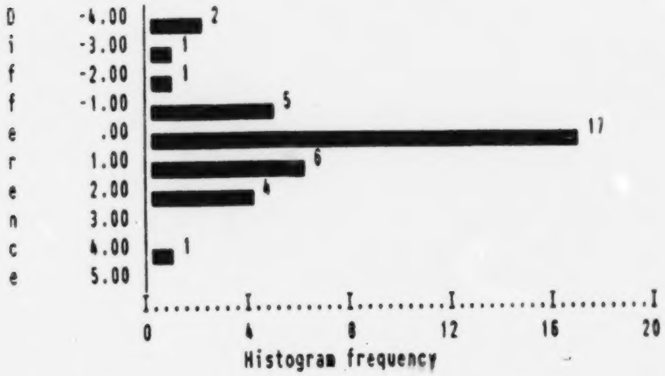


Figure 3. Distribution of difference in rank order of state mean based on overall mathematics composite and state mean based on the geometry subscore

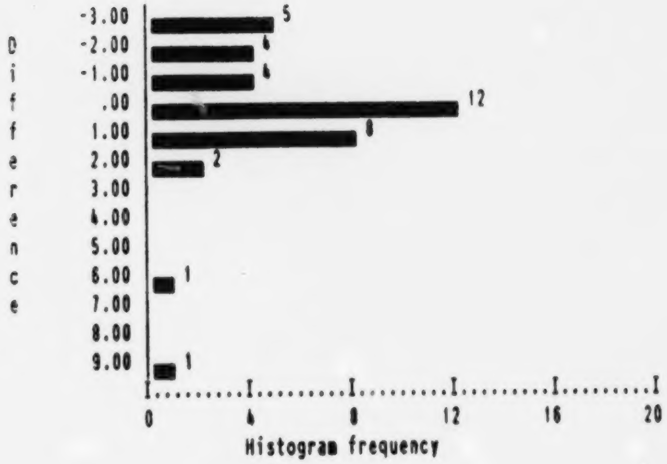


Figure 4. Distribution of difference in rank order of state mean based on overall mathematics composite and state mean based on the data analysis, statistics, and probability subscore

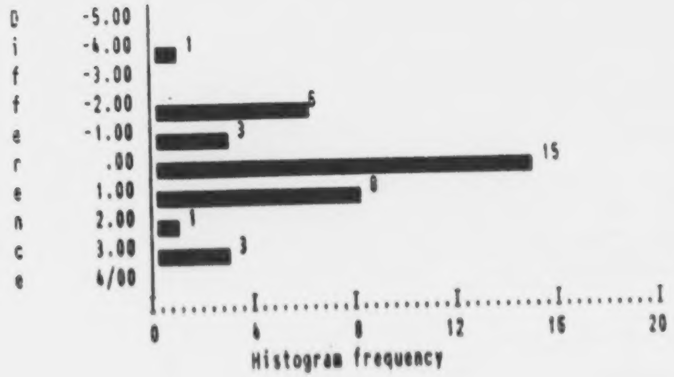
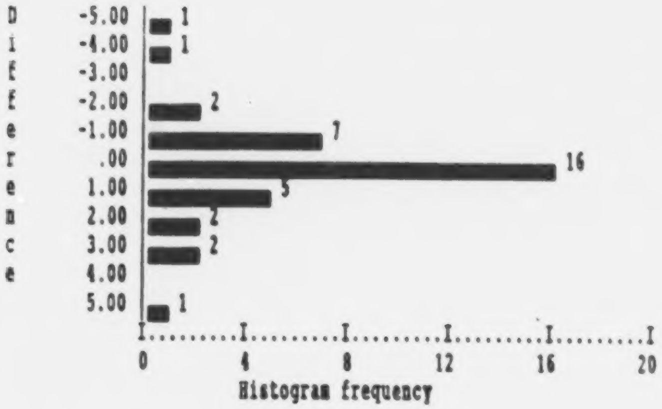


Figure 5. Distribution of difference in rank order of state mean based on overall mathematics composite and state mean based on the algebra and functions subscore



The categorization of items by Silver (1991) according to the themes in the NCTM standards (problem solving, communications, reasoning, and the remainder of the items that were classified as fitting none of the NCTM themes) also produced subscores averages that ranked states in much the same way as the average on the overall composite. (One NCTM theme, connections, used by Silver was not included in these analyses since only one item was placed in that category). As is shown in Table 4, the rank order correlation of the average overall composite score with each of the average subscores based on the NCTM themes was again .99 for all four subscores. None of the rank orders based on one of the subscale averages differed from that based on the average overall composite by more than 5 ranks. All but 6 of the 148 differences in rank orders of the four subscale averages from the overall composite average were 3 ranks or less. As in the case of content categories, the state comparisons in terms of average performance are not very sensitive to either the NAEP process categories or the use of the themes in the NCTM standards to classify items for the calculation of subscores.

Percent correct scores were computed separately for the 102 multiple-choice items and the 35 open-ended items that required students to provide a numerical or written response. The rank-order correlation between the multiple-choice and open-ended average percents correct was .95. The distribution of differences in rank order for scores based on multiple-choice and open-ended items is shown in Figure 6. As can be seen, for 4 of the 37 states the rank-order would differ by 6 or more ranks depending on whether the ranking was determined by average percent correct on multiple-choice or open-ended items.

Although state comparisons appear to be sensitive to the choice of item format, it should be noted that the open-ended subtest results are less stable than the multiple-choice subtest because there are fewer open-ended items included in the assessment. Nonetheless, the change of 11 ranks that was observed for one state is substantial and could lead to rather different conclusions about the relative performance of students in that state. Given the current push to increase the use of open-ended items while reducing the emphasis on multiple-choice items, it will be important in future assessments to consider the degree to which the two types of assessment exercises convey similar or discrepant messages about the achievement of students in different states.

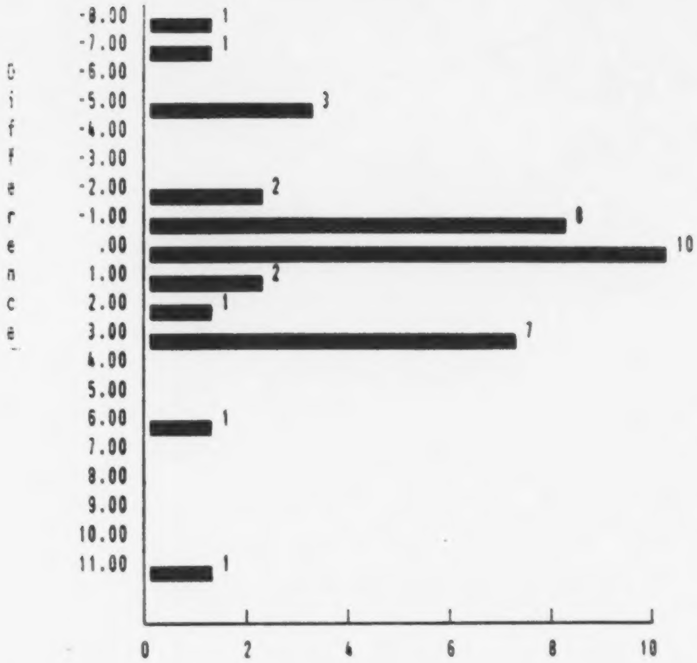
Table 4. Intercorrelations among state ranks based on mean overall composite scores and each of the four subscores based on themes from the NCTM standards*

Variable	Overall Composite	Problem Solving	Communication	Reasoning	None of the NCTM Themes
Overall Composite	1.00	.99	.99	.99	.99
Problem Solving	.99	1.00	.98	.98	.98
Communication	.99	.98	1.00	.98	.98
Reasoning	.99	.98	.98	1.00	.98
None of the NCTM Themes	.99	.98	.98	.98	1.00

N = 37 States.

* Categorization of items for themes from NCTM Standards Based on Silver (1991). One theme, connections, was excluded because only one item was categorized in that category. The number of items in the remaining categories was Problem Solving, 19, Reasoning 41, Communication 12, and None of the NCTM Themes, 64.

Figure 6. Distribution of difference in state rank order based on open-ended items and rank order based on multiple-choice items



Metric

Simple weighted average percent correct scores for each state were compared to the more complex state averages of overall mathematics composite scores. The rank order correlation between the two average scores for states was .997. None of the state ranks would change by more than 2 if the simple weighted average percent correct scores were used in place of the current NAEP average composite scores. Clearly, the change in metric would have no practical implications for comparison of state average scores.

Alternative Statistical Summaries

State means provide only limited information about the overall distribution of performance of students. They do not tell how the highest or lowest achieving students within a state perform relative to their counterparts in other states. In order to determine the effect of considering other parts of the score distribution on state-by-state comparisons, selected within-state percentile points and the percentage of students scoring about the NAEP anchor points of 200, 250, and 300 were analyzed and compared to the use of state averages.

The rank-order correlations of state means and percentile ranks of 10, 25, 75, and 90 are listed in Table 5. As might be expected, the relationships of the rank order based on means with those based on the 25th or 75th percentile are stronger than they are with the rank order based on either of the two more extreme percentiles. Although the rank order correlations between state means and the 10th or 90th percentile (.94 and .93, respectively) are still substantial, they reflect some sizeable shifts in the rank order of individual states.

The overall composite state medians, 10th percentiles, and 90th percentiles are plotted in Figure 7 with the states ordered from high to low based on the state medians. As can be seen from the jaggedness and absence of rank order from left to right for the points representing the 10th and 90th percentiles, the relative standing of a state can be quite different depending on whether the focus is on the median or the performance of high or low achieving students within each state.

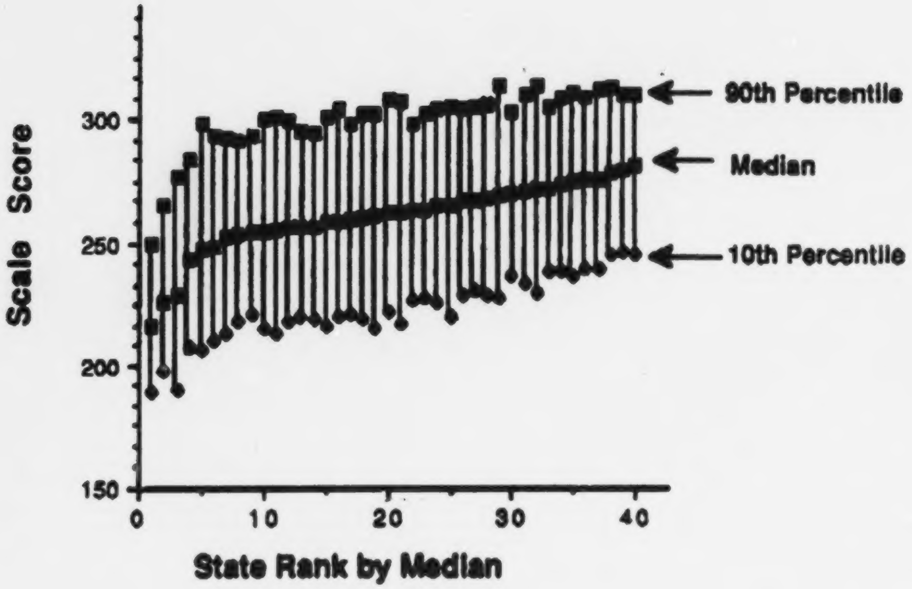
The distribution of differences between ranks based on means and ranks based on the 10th and 90th percentiles are shown in Figures 8 and 9, respectively.

Table 5. Intercorrelations among state ranks based on mean overall composite scores and selected within state percentile ranks of student scores on the overall composite

		Percentile Rank			
Variable	Mean	10th	25th	75th	90th
Mean	1.00	.95	.98	.98	.92
10th	.95	1.00	.98	.88	.79
25th	.98	.98	1.00	.93	.85
75th	.98	.88	.93	1.00	.97
90th	.92	.79	.85	.97	1.00

N = 37 States.

Figure 7. Plot of state medians, tenth percentiles, and ninetieth percentiles with states ordered according to means



As can be seen in Figure 8, five states would have a shift in rank of seven or more by moving from the mean to the 10th percentile. Maryland and New York would have their rank orders decrease by seven or more if the states were ordered on the 10th percentile rather than on the mean. Kentucky, New Mexico, and West Virginia, on the other hand, would move up by seven or more ranks by switching from the mean to the 10th percentile.

Moving from the mean to the 90th percentile (Figure 9) would lead to shifts of six or more ranks for seven states. Four states, Connecticut, Maryland, New Jersey, and Virginia, would show increases of six or more ranks, while three other states, Arkansas, Indiana and Oklahoma, would show decreases of six or more ranks if the 90th percentile rather than the mean were used to order states.

The change in rank order based on the 10th or 90th percentiles in comparison to means reflects, of course, differences in the within-state variability of performance. This "explanation" of the change, however, only underscores the need to consider more than central tendency when comparing states. The fact that the 90th percentile is 95 points higher than the 10th percentile in Hawaii, while the corresponding difference is 66 points in Montana conveys comparative information about the mathematics performance of students in those to states that is not revealed by a comparison of their means.

NAEP has traditionally reported achievement in terms of the percentage of students scoring above anchor points of 150, 200, 250, 300, and 350. On the cross-grade scale, the 150 and 350 anchor points are too extreme to be useful in reporting grade 8 mathematics results: however, the other three anchor points are potentially helpful. The rank order correlations of the overall mathematics composite and the percentage of students within a state scoring above 200, 250, and 300 are shown in Table 6. Although the correlations of ranks based on the mean and those based on the percentage of students above 250 is quite high, the correlations between these mid-scale rankings and rankings based on the percentages above 200 and 300 reflect some sizeable shifts in rank. The correlation of ranks for means and those for the percentage above 200 is comparable to those found for the correlation of ranks for means with those for the 10th and 90th percentiles, and as would be expected, there are similar shifts in rank orders. Florida, for example, would have its rank increase by four by shifting from means to the percent of students scoring above 300, while the same shift would lead to a decrease of nine ranks for Oklahoma.

Table 6. Intercorrelations among state ranks based on mean overall composite scores and the percentage of students scoring above the NAEP anchor scores of 200, 250, and 300

Variable	Mean	Percent Above 200	Percent Above 250	Percent Above 300
Mean	1.00	.93	.99	.96
% Above 200	.93	1.00	.93	.84
% Above 250	.99	.93	1.00	.96
% Above 300	.96	.84	.96	1.00

N = 37 States.

Figure 8. Distribution of difference in rank order of overall mathematics composite state mean and tenth percentile

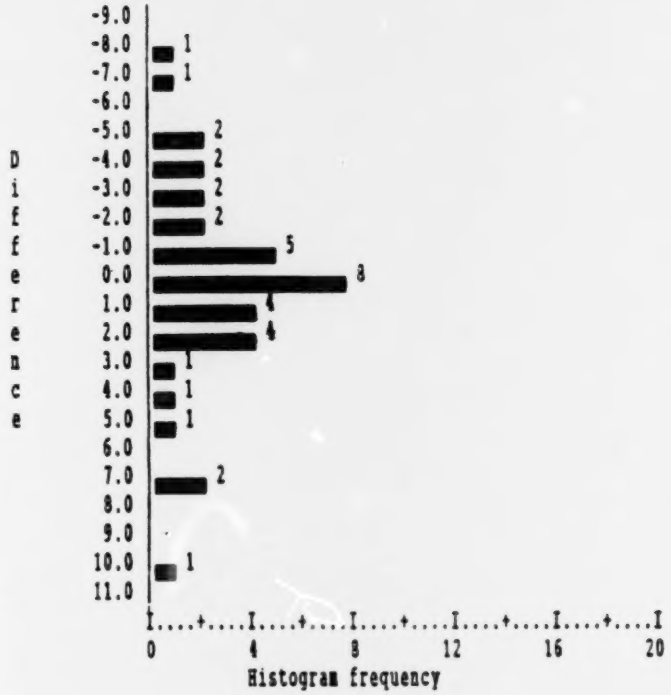
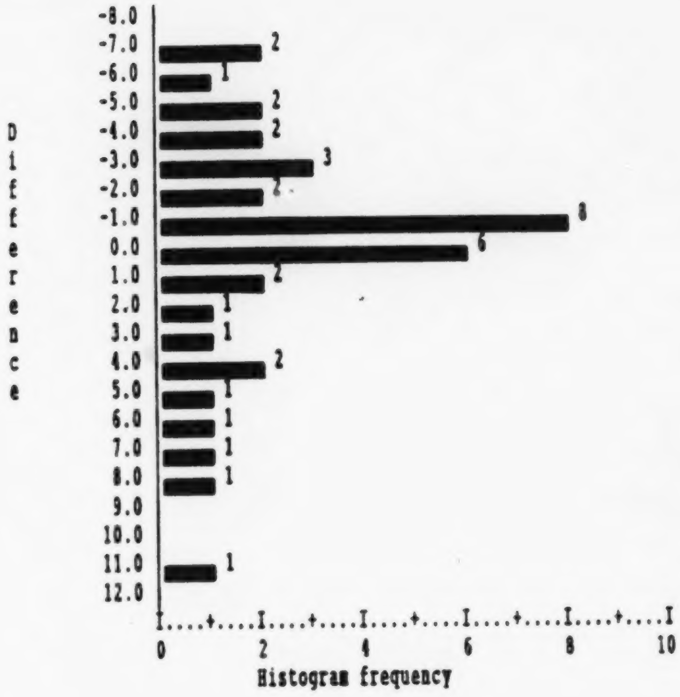


Figure 9. Distribution of difference in rank order of overalll mathematics composite state mean and ninetieth percentile



Subgroup Analyses

Many subpopulations could be defined within each state and used as the basis for comparisons among states. Only two ways of defining subpopulations were investigated here. These are racial-ethnic groups and samples defined by students attending the top and bottom third of the schools within each state.

State averages for the total sample on the overall mathematics composite were compared to state averages based on the state samples of white students, black students, and Hispanic students where the sample sizes were judged by NAEP to be adequate to produce reliable estimates (Mullis, et al. 1991). All 37 states had sufficient samples of white and Hispanic students to produce estimates. However, only 28 of the 37 states had sufficiently large samples of black students for estimates to be reported in the TSA report.

The rank order correlations of total sample averages, and the averages for white, black, and Hispanic students are shown in Table 7. In general, these correlations are smaller than those that were reported for previous analyses. Even the correlation between the total sample averages and the white student averages is only .85. The correlation between the total sample averages and the Hispanic student averages (.51) is much lower and reflects considerable differences in the standing of states in terms of average performance for all students and the average performance for Hispanic students.

The rank orders of the average mathematics achievement of white, black, and Hispanic students can be quite different. The relatively large shifts in rank order for these three groups of students is evident from Figures 10 and 11. Figure 10 displays the differences in rank order for white students and that for Hispanic students, while Figure 11 displays the corresponding differences for white and black students. As can be seen, some of the changes in rank order are quite large.

Another indication of the magnitude of shifts in rank order when the performance of students in different racial-ethnic groups is used to compare states is provided by dividing the states into quartiles. Tables 8 and 9 display cross tabulations of the quartiles based on rankings for white students compared with black students and for white students compared with Hispanic students, respectively.

Table 7. Intercorrelations among state ranks based on mean overall composite scores for the total sample and for state samples of white, black, and Hispanic students

Variable	Total Sample	White Students	Black Students*	Hispanic Students
Total Sample	1.00	.85	.76	.51
White Students	.85	1.00	.56	.49
Black Students	.76	.56	1.00	.37
Hispanic Students	.51	.48	.37	1.00

* Correlations involving black student averages are based on 28 of the 37 states since 9 states had samples of black students that were too small to produce reliable estimates. The remaining correlations are based on 37 states.

Figure 10. Distribution of difference in rank order of state mean based on overall mathematics composite for white students and state mean based on Hispanic students

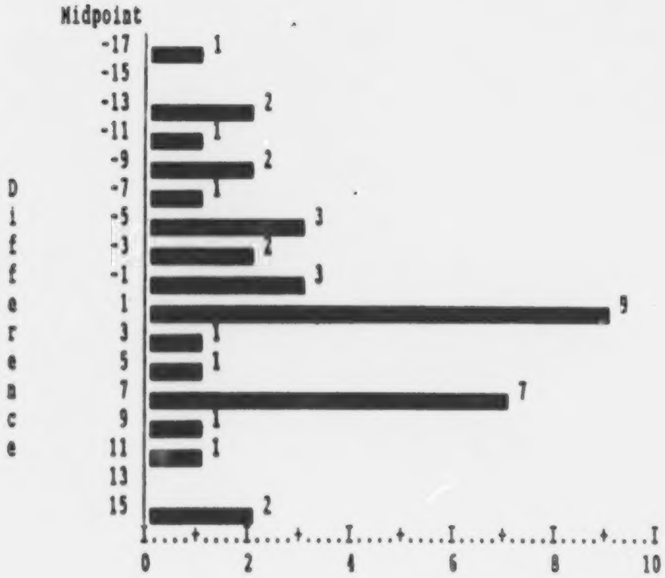
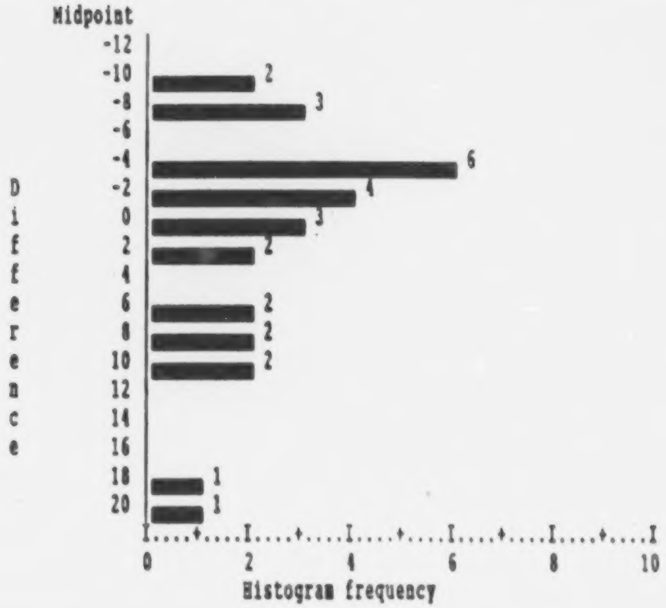


Figure 11. Distribution of difference in rank order of state mean based on overall mathematics composite for white students and state mean based on black students with sufficient sample size for reliable estimates*



* In nine states the sample size for black students was insufficient to produce reliable estimate according to the TSA report (Mullis 1991). Hence, these rank order differences are based on 28 states.

As can be seen in Table 8, only two of the seven states that rank in the top quartile based on the performance of black students also rank in the top quartile based on the performance of white students. One state that is in the top quartile for black student performance is in the lowest quartile based on white student performance. Large shifts in quartiles are also seen in Table 9 where quartiles based on the achievement of Hispanic students are compared with those based on the achievement of white students. It is evident that a number of states that rank relatively high in terms of the performance of students in one racial-ethnic group rank relatively low in terms of performance of students in another racial-ethnic group.

The difference in means on the overall mathematics composite score for white and black students ranged from a low of 20 points on the NAEP scale in Kentucky to a high of 43 in Wisconsin. The percentage of students who are black is approximately the same in these two states (9 percent and 8 percent for Kentucky and Wisconsin, respectively). A similar range of differences in means between white and Hispanic students was observed. Florida, where 17 percent of the students were Hispanic, had a mean difference of 19 points on the NAEP scale while Connecticut, where 10 percent of the students were Hispanic, and Pennsylvania, where 5 percent of the students were Hispanic, each had a mean difference of 42 points.

Obviously, there are many possible reasons for the group differences being twice as large in one state as in the other, including migration patterns, concentrations in urban areas, differential poverty rates, and identifiable subpopulations within the larger groups being compared. The marked contrasts between states in the magnitude of the group difference in average achievement also raise questions, however, about differential access to education.

One indication of differential access comes from the NAEP categorization of schools in each of the states according the average overall mathematics composite scores of students within each school. Performance results for students in schools in the top and bottom thirds within each state were included in the TSA report (Mullis et al. 1991). By definition, students in the top third of the schools will have higher average performance than students in the bottom third. The magnitude of those differences, however, depends on the degree to which students have differential access to schools on the basis of characteristics related to their mathematics achievement.

Table 8. Cross tab of state mean overall composite score by quartiles based on black and white student achievement within a state

State Standing Based on White Student Achievement	State Standing Based on Black Student Achievement				
	LOWEST QUARTER 1.00	SECOND QUARTER 2.00	THIRD QUARTER 3.00	HIGHEST QUARTER 4.00	Total
1.00 LOWEST QUARTER	5	1		1	7
2.00 SECOND QUARTER	2	4		1	7
3.00 THIRD QUARTER		2	2	3	7
4.00 HIGHEST QUARTER			5	2	7
Column	7	7	7	7	28

Table 9. Cross tab of state mean overall composite score by quartiles based on Hispanic and white student achievement within a state

State Standing Based on White Student Achievement	State Standing Based on Hispanic Student Achievement LOWEST QUARTER 1.00	SECOND QUARTER 2.00	THIRD QUARTER 3.00	HIGHEST QUARTER 4.00	Total
1.00 LOWEST QUARTER	8		2		10
2.00 SECOND QUARTER	1	5	3		9
3.00 THIRD QUARTER	1	2	3	3	9
4.00 HIGHEST QUARTER		2	1	6	9
Column	10	9	9	9	37

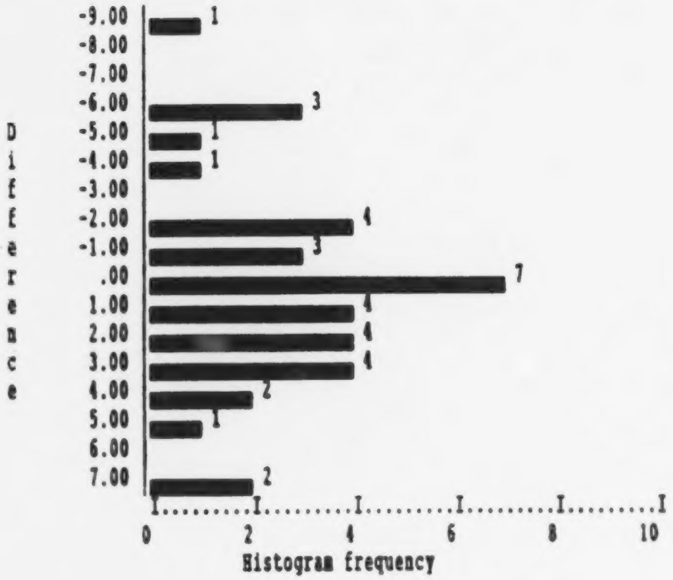
Rank-order correlations of average overall mathematics composite scores for the total sample and for students in the top and bottom third of schools within a state are shown in Table 10. Differences in rank order of states using results for students in the top third and the bottom third of schools are shown in Figure 12. Although there is a substantial correlation between the averages for students in the top and bottom third of schools within a state, five of the states would have their rank change by six or more depending on whether ranks were based on students in the top or bottom third of schools within a state.

Table 10. Intercorrelations among state ranks based on mean overall composite scores for the total sample and for the samples of students attending the top and bottom one-third of schools within the state

Variable	Total Sample	Top Third of Schools	Bottom Third of Schools
Total Sample	1.00	.93	.99
Top Third of Schools	.93	1.00	.94
Bottom Third of Schools	.99	.94	1.00

N = 37 States.

Figure 12. Distribution of difference in rank order of state mean based on overall mathematics composite for sample of students attending top one-third of schools and state mean based on students attending bottom one-third of schools



Conclusions and Recommendations

The relative standing of a state was altered relatively little for reporting metrics including the NAEP scaled composite scores, scaled content subscores, process subscores based on either the NAEP framework or the NCTM themes, or simple percent correct scores. Several other factors that were investigated had substantial effects on the relative standing of a state, however.

States that have high or low standing in terms of average student achievement may rank quite differently in terms of the performance of their highest or lowest achieving students. States where one subgroup of students compares quite favorably to their counterparts in other states, sometimes compare quite unfavorably to other states in terms of the achievement of students in another subgroup. Such differences raise important questions about differential access to educational opportunities and deserve greater attention.

Greater attention also needs to given to characteristics of the distribution other than the mean. The TSA report appropriately highlights more comprehensive information in figures showing state distributions and in extensive tables of results (see, for example, Figure 11.2 in Mullis et al. 1991, 239). Unfortunately, the press accounts of the TSA results have given little attention to such information.

The results of our analyses underscore the importance of going beyond simplistic comparisons based on average overall achievement of all students within a state. The full distribution of scores demands greater attention than it has generally received. Equally important is the need to consider the achievement of critical subpopulations of students.

References

Cronbach, L. J. 1971. "Test Validation," in *Educational Measurement* (2nd ed.) Ed. R.L. Thorndike. Washington, DC: American Council on Education.

Mazzeo, J. 1991. "Data Analysis and Scaling." in *The Technical Report of NAEP's 1990 Trial State Assessment Program.* Ed. S. L. Koffler et al. Report No. 21-ST01. Washington, DC: National Center for Educational Statistics.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G. W. 1991. *The State of Mathematics Achievement: NAEP's 1990 Assessment of the Nation and the Trial Assessment of States.* Washington, DC: National Center for Educational Statistics.

Novick, M. R. 1982. "Educational Testing: Inferences in Relevant Subpopulations." *Educational Researcher,* 11, 8: 4-10, 28.

Spencer, B. D. 1983. "On Interpreting Test Scores as Social Indicators: Statistical Considerations." *Journal of Educational Measurement,* 20: 317-333.

Walker, D. F., and Shaffarzick, J. 1974. "Comparing Curricula." *Review of Educational Research,* 44: 83-111.

A Study of the Impact of Reporting the Results of the 1990 Trial State Assessment: First Report

Frances B. Stancavage
American Institutes for Research

Edward Roeber
Council of Chief State School Officers

George H. Borhnstedt
American Institutes for Research

January 1992

Purpose of the Study

On June 6, 1991, a long-awaited event occurred. The first state-by-state results from the 1990 National Assessment of Educational Progress (NAEP) Trial State Assessment (TSA) of eighth grade mathematics were reported to the nation. Because of its potential consequences, this event, more than any other connected with this program, was the subject of controversy. Would the reporting of the NAEP TSA cause local districts and states to change the *curriculum* or *instruction* that is provided to students? Would local or state *testing programs* change to accommodate NAEP-tested skills, would they remain as they are, or would they simply be pushed aside? Would any such changes in curriculum or assessment, should they occur, be judged as positive by mathematics educators and others, or would the changes be viewed as regressive and counter-productive? Finally, would it be found that the entire NAEP TSA effort had no impact at all and was, therefore, a wasteful expenditure of time and money?

The purpose of this study, as one part of an extensive and on-going evaluation of the NAEP TSA authorized by Congress in 1988, is to gather information about the impact of the reporting of the results from the 1990 TSA. While other studies will investigate such features as the quality of the assessment frameworks and items used in the TSA, the sampling methodology, the data collection procedures, and the standard-setting practices, this study is intended to answer questions about how the results were reported, and the nature of the impacts, if any, arising from the reporting of such results.

In this, the first report of the impact of reporting study, preliminary answers are provided to the following questions.

- How did the individual states plan for, and then handle, the June 1991 reporting of TSA results? What plans were they making, as of July or August, to respond to the September release of achievement levels?
- How useful were the reports and other materials issued by NCES? What suggestions do the respondents have for improving these reports?
- How were the results of the TSA reported by the press?
- How do the respondents evaluate the press coverage received by the TSA?
- How widely is NAEP known? By July-August 1991, how many of the potential consumers of TSA data had seen the results?

- How highly do different respondent groups value the NAEP TSA as a monitor of education programs? What do they see as NAEP's major strengths and weaknesses?

- Do the respondents think that the NAEP state assessments should be continued after 1992? Would they like to see the NAEP program expanded to make results available for individual districts, schools, or students?

- Have the 1990 TSA results given rise to meaningful discussions among educators and policymakers at the national, state, or local levels?

- What evidence is there that the TSA is beginning to influence state mathematics instruction or assessment?

- Overall, was the 1990 NAEP TSA considered a useful exercise?

As more time passes since the reporting of the 1990 TSA, and data collection for this impact of reporting study continues, answers to these questions will be elaborated and refined. In addition, the first experience of using the achievement levels in reporting (in September 1991) will be explored in depth, with respondents asked to compare the clarity and the utility of this reporting device as compared to the anchor levels used in June 1991. Finally, efforts will be made to determine the contextual factors that appear to influence the nature and extent of the impact on education in the various states attributable to the TSA.

Methodology

The study, which is on-going, is collecting data from a number of sources. First and foremost among the data sources are periodic telephone interviews with samples of national, state, and local educators and policymakers most concerned about the direction of mathematics education and best placed to influence the process. In spring 1992, following the third cycle of telephone interviews, written responses to a series of open-end questions will also be elicited from knowledgeable individuals in selected localities to try to identify, in the greater detail of a case-study format, the contextual factors behind any education impacts that are observed. Finally, various artifacts of the reporting process, such as press releases and news articles, are being examined for completeness and accuracy in order to provide an analysis of the types of information that were reported, as well as the manner in which they were reported.

Surveys on Reporting

Respondent samples. It was recognized from the outset that many different people would have opinions about this important assessment project and the reporting of results from it, but both logistics and budget required that some prioritization of groups and individuals be made. Selection was therefore made from those groups and individuals who were most concerned about the reporting and those most in a position to observe the impacts of it.

Table 1 shows the various groups and individuals being interviewed and the number of respondents identified in each respondent category.

Table 1. Types and numbers of respondents sampled

Level	Group	Sites	Respondents/ site	Total
National educators	NAEP Math Planning Committee			11
	NAEP Math Steering Committee			14
	Current presidents--national education organization			14
National policymakers	House and Senate staff			24
All states	Mathematics specialists	51	1	51
	Assessment directors	51	1	51
State sample	State politicians	10	2 or 3	25
	Chief State School Officers	10	10	10
	Presidents--state education organizations (PTA, teachers' unions, teachers of mathematics)	10	2 to 9	48
Local	Building principals	30	1 to 3	84*
	District superintendents	30	1	30
	District curriculum and assessment specialists	30	0 to 2	34*

* Numbers from first interview cycle. Will vary slightly in subsequent cycles when new samples are drawn.

National respondents include all education-oriented Congressional staffers serving committee members on relevant House and Senate committees, as well as key individuals who were involved in the planning and consensus process for the 1990 NAEP TSA. Congressional

BEST COPY AVAILABLE

staffers were drawn from the Senate Labor and Human Resources Committee, the House Education and Labor Committee, and the Education and Labor, Health and Human Services subcommittees of the Senate and House Appropriations Committees.

The NAEP planning and consensus process is represented in the sample by all members of the two committees convened by the Council of Chief State School Officers (CCSSO) to oversee the development of the 1990 NAEP mathematics specifications. The first of these committees, the Steering Committee, was composed of representatives of various national education organizations such as the National Education Association, the American Federation of Teachers, the National Association of Secondary School Principals, and the American Association of School Administrators. The second committee, the Planning Committee, included those mathematics and assessment experts who actually wrote the mathematics assessment framework and developed initial blueprints for the assessment. In addition, interviews are being conducted with the current presidents or directors of all education organizations that sent representatives to the Steering Committee.

State-level respondents include the state assessment directors and mathematics specialists in each of the 50 states and the District of Columbia. Additional samples of potential state- and local-level end users of TSA data were drawn from 10 sample states--8 that chose to participate in the TSA and 2 that chose not to participate. In designing these latter samples, the goal of maximizing diversity in the overall respondent pool was balanced against the goal of obtaining sufficient respondents in the case study states to help in characterizing the impact of reporting within a given state context. Moreover, the number of respondents was traded off against the value to be gained from using in-depth telephone interviews rather than simpler mail surveys.

To draw the state sample, participating states were stratified by dividing them into the four NAEP regions and then dichotomizing them by high and low median NAEP test scores within region. Nonparticipating states were dichotomized by population size. One state was sampled from each of the resultant sampling cells. State assessment directors and mathematics specialists were called upon to help identify appropriate education organizations in the sampled states; representatives of these organizations were then added to the respondent pool, along with the Chief State School Officers and education-oriented state legislators.

For the local sample, 3 local education areas (LEAs) were sampled randomly within each of the 10 sample states. Respondents included the district superintendent, a random sample of building principals, and district assessment and curriculum specialists, where present. In order to avoid any response bias that might arise from sensitizing individuals at the periphery of the

NAEP process via earlier questioning, a new set of LEAs is being drawn for each data collection cycle. (At the state and national levels, the same individuals are being interviewed at all data points. This will allow the responses of these individuals to be tracked across time.)

Data collection schedule. TSA results for 1990 were first released by the National Center for Education Statistics (NCES) in early June 1991; a second data release occurred in late September 1991, when the National Assessment Governing Board (NAGB) and the National Education Goals Panel (NEGP) both presented 1990 TSA results using variants of the NAGB achievement levels. In light of this release schedule, four survey dates were selected for the telephone interviews: (1) April-May 1991; (2) July-August 1991; (3) January-February 1992; and (4) October-November 1993. These dates were chosen because they were: (1) just before the June release of results and while the states were in the midst of planning their own responses to the release; (2) a few weeks following the release, when the most immediate consequences would be apparent but the details of the release process would not yet be forgotten; (3) seven months after the June release (three months after the release of the NAEP data according to the NAGB levels); and (4) 16 months after the release, during the following school year, when longer-term consequences might be expected to become evident. (The timing of the fourth interview cycle was also influenced by the release dates for the first 1992 TSA results. The interviews will occur just prior to that release and will be used to capture planning for that release cycle as well as evidence of longer term consequences from the 1990 TSA.) Table 2 shows an overview of the schedule for the survey interviews.

Table 2. Overview of interview schedule

	Respondent group				
	National educators	National policy-makers	All states	State sample	Local
April-May 1991	X	X	X	X	X
July-August 1991	X	X	X	X	
January-February 1992	X	X	X	X	X
October-November 1993	X	X	X	X	X

No local sample was drawn for the July-August 1991 data cycle because of the difficulty of locating such respondents during the summer holiday period.

Survey instruments. Survey questions covered two broad areas: the nature of the reporting process as it occurred in each state, and the impact of reporting. The primary respondents for questions on the reporting process itself were the mathematics specialists and assessment directors in the 50 states and the District of Columbia. In the design of this part of the survey, an underlying hypothesis was that careful and timely planning is necessary in order to appropriately manage the impact of reporting. Questions were therefore framed to collect information about the planning process as well as other aspects of the reporting.

Questions aimed at determining the *impact* of the reporting included the following.

- the perceived adequacy of the reports and other materials provided by NCES to aid in dissemination

- the way in which the data were picked up and used (if at all) by the press, policymakers, business leaders and others

- the respondent's evaluation of the potential or actual value of NAEP results, the adequacy of the reporting process in this cycle, and the contextual factors that affected the impact of reporting (positive or negative)

- increases in knowledge/or understanding--whether the reporting of Trial State NAEP results increased knowledge about how student scored on NAEP, what NAEP is and what it is intended to do, what the structure and content of math education is and how it relates to the activities of daily living. Also, within a given state, whether respondents were more knowledgeable post-release about the specific messages that the state representative sent out to convey in his or her reporting of results.

- changes in attitudes--whether the reporting modified attitudes about NAEP (e.g., NAEP does/or does not provide a useful basis for comparison between states; NAEP should/or should not be used for district or school level comparisons); standardized testing in general, particularly as a tool for assessing the education system; education and, specifically, math education; levels of student achievement, particularly in math.

- intentions and actions--reported actions or plans as a consequence of the reporting: changes in future participation in Trial State NAEP, in reporting of Trial State NAEP, or both; changes in state testing programs; changes in curriculum or instruction, particularly math.

Response rates. Overall response rates for the first two data cycles have been good, but response rates have varied by respondent group. Table 3 shows the numbers of completed interviews and the response rate by respondent group and data cycle.

Table 3. Numbers of completed interviews and response rates by respondent group and data cycle

	April-May		July-August	
	Completed interviews	Response rate	Completed interviews	Response rate
National educators	23*	59%	29*	74%
National policymakers	19	79%	18	75%
Assessment directors and mathematics specialists	96	94%	102	100%
State sample	54	65%	74	89%
Local sample	110	74%	NA	NA
TOTAL SAMPLE	302	76%	248	90%

* Includes four Planning and Steering Committee members who were interviewed in their capacity as state assessment directors or state mathematics specialists.

Case Study Analysis

As the interview work proceeds across the points of time identified, an attempt will be made to document what has happened in the eight states selected in the state sample, as well as in instances of particularly good, or particularly bad, uses of NAEP data at the state or local levels. Overall, no more than 15 cases will be identified (8 states in the state sample plus 7 special cases). These latter cases will be identified either through the second or third interviews or through an examination of the various press reports about the 1990 NAEP TSA. The actual case studies will be carried out after the third round of interviews and will be documented in a later report.

For each identified case, one individual will be contacted and asked to complete a written, structured open-end survey in order to describe (from his or her perspective) the activities that occurred over the span of time in which the interviews took place (i.e., before, during, and after the NAEP TSA reporting) and the context in which these activities occurred. Then, each respondent will be interviewed by telephone by one of the principal investigators or a representative of American Institutes for Research. From both the written description and telephone interview, the principal investigators will compile short, written case studies that will eventually be made part of the overall study report.

In the case studies, an attempt will be made to go beyond simple description to determine the factors underlying the incident. While this is somewhat problematic due to the post-hoc basis on which information will be collected and the impressionistic types of information being used, it is important information to attempt to ascertain because it could potentially provide valuable feedback on the mechanisms that can be used to improve the reporting of future NAEP state assessments.

Analysis of Reporting Artifacts

A national clipping service was retained from June 1991 through December 1991 to compile NAEP-related news articles appearing in major national papers or in any newspapers in the 10 sample states. In addition, project staff attended the national press conferences in June and September 1991 at which NAEP data were presented, and survey respondents were asked to forward state press releases and articles written locally about NAEP.

The purpose of this part of the investigation is to document, first hand, the ways in which the results were released to the public, to determine the volume of national and local media coverage generated by the release, and to analyze, at least in general terms, the content and tone of the various pieces. Specific attention is also being paid to the portrayal of state comparisons or rankings and the ways in which the statistical complexities of these comparisons are being handled by the media and others.

This report includes a discussion of the national media coverage of the June 1991 data release. Coverage at the state and local levels, and coverage of the second, September, release will be summarized in a later report.

Research Findings

This section contains preliminary summaries of results from the first two interview cycles and an analysis of June 1991 press clippings from the 50 largest U.S. papers by circulation. The next report, which will be available by summer 1992, will encompass results from the third interview cycle, the case-study analyses, and a more comprehensive analysis of reporting artifacts. Longitudinal analyses of data from the three interview cycles, including analyses of changes in knowledge and attitudes over time for individual respondents, will also be covered at that time.

The results in this section are organized around a series of questions, listed in the introduction to this paper, that helped to guide the analyses. The text of the survey questionnaires, as well as complete response frequencies for all respondent types, are available from the authors.

How did the individual states plan for, and then handle, the June reporting of TSA results? What plans were they making, as of July or August, to respond to the September release of achievement levels?

As stated earlier, April-May was selected for the first cycle of interviews in part because the investigators felt that it was close enough to the release date to allow for the capture of any significant planning activities that might occur. The interviews with state assessment directors and mathematics specialists showed that over 60 percent of the states had held internal meetings to discuss the NAEP TSA program and its results, while a sizable minority (43 percent) had also identified communication goals for the release, begun to draft a press release, or taken some other more explicit action towards planning for the release.[1]

In July-August, these respondents reported that all but one of the states that had participated in the 1990 TSA had prepared a press release for the June release of results, and about two thirds had held a press conference. Personal briefings, typically of the governor, the state superintendent, and members of the state board of education, also occurred in about two thirds of these states, while respondents in over a quarter of these states wrote articles or gave speeches to disseminate the results further. Dissemination activities in states that had *not* participated in the 1990 TSA were much more limited, with only three states issuing press releases and one holding a personal briefing. (See Table 4.)

---

[1] For descriptive information about the planning and dissemination activities carried out by the state, responses from the state assessment director and the state mathematics specialist were cumulated. That is, if either one reported that a particular activity occurred, it was presumed to have occurred, even if the other did not mention it.

Table 4. State reporting activities for the results of the 1990 TSA
June 1991

Note: for this table the District of Columbia is counted the same as a state.

	Participating states (N=38)	Nonparticipating states (N=13)	Total (N=51)
Press release	37 (97%)	3 (23%)	40 (78%)
Press conference	24 (65%)	0 (0%)	24 (47%)
Personal briefings	25 (68%)	1 (8%)	26 (51%)
Speeches or articles	10 (27%)	0 (0%)	10 (20%)

The actual presentation of the TSA results was handled most often by the state superintendent, assessment director, and mathematics specialist. The messages emphasized by the states were the need to improve instruction, the fact that the state is doing well but reform is still needed, and the level of state performance and the strengths and weaknesses of its performance.

Looking back, most (90 percent) of the state mathematics specialists and assessment directors (90 percent) said they felt very well or well prepared for the release. In fact, over 40 percent said that there was nothing more that NCES could have done to better prepare them for the release of the results. However, some respondents felt that NCES should have done more, including providing press releases and related materials, providing more materials that could be used with lay public to clarify what NAEP data can and cannot show, and, generally, undertaking additional prerelease communications with the states about the results and the report formats.

As of July-August, two thirds of the states reported ongoing activities related to the June 6 release, including organizing groups to address the problems uncovered, holding conferences to disseminate the results, distributing the NCES data reports to various groups and individuals, and conducting their own analyses of the NAEP data.

When asked what actions they were taking to prepare for the September 30 re-release of the data by NAGB, only a few of the states reported having begun any formal planning for the event or having any clear idea of how they would present these data to the press. Nearly a third of the states that had participated in the TSA indicated that they were unlikely to do anything to publicize or comment on these achievement level data, and another sizable proportion were unsure what they might do. In most cases, the decision not to publicize seemed to arise from a feeling that one airing was enough for the 1990 TSA results. A few respondents, however, noted that they could not really plan a meaningful response because NAGB was not giving them any advance notice of what the data would look like, and a few others indicated that they would rather not get involved with the achievement levels until it became clear whether they would be retained in their present form.

How useful were the reports and other materials issued by NCES? What suggestions do the respondents have for improving these reports?

As shown in Table 5, among states that participated in the 1990 TSA, all but one used the separate NCES report for their own state in their reporting process. High proportions also reported using the composite report, the executive summary, the NCES press kit, and the national press conference closed-circuit television feed. Among states that did *not* participate in 1990, about half nevertheless reported using the composite report, and a similar proportion reported using the executive summary.

Table 5. State use of NCES reports and press materials, June 1991

Note: for this table the District of Columbia is counted the same as a state.

	Participating states (N=38)	Nonparticipating states (N=13)	Total (N=51)
State report	37 (97%)	NA	37 (73%)
Composite report	36 (95%)	6 (46%)	42 (82%)
Executive summary	34 (89%)	7 (54%)	41 (80%)
Press kit	30 (79%)	0 (0%)	30 (59%)
Closed-circuit television feed from national press conference	28 (74%)	0 (0%)	28 (55%)

Despite these high use rates, nearly half of all the state assessment directors and mathematics specialists indicated that they experienced some problems with the materials. The most commonly cited concerns were the volume of the data (13 percent), the complexity of the data (14 percent), and the references to grade-level equivalents (15 percent). These respondents provided a wide variety of suggestions for ways in which the reports and the reporting process could be improved in the future, including (1) producing smaller, simpler reports for specific audiences, such as parents or legislators; (2) finding better ways to highlight comparisons between states with similar demographics; and (3) involving other organizations, such as the National Council of Teachers of Mathematics (NCTM), in the reporting and interpretation of results.

The majority of the state assessment directors and mathematics specialists felt that each of the various types of information provided in the NCES reports was useful to them. Levels of performance in the separate math content areas got the highest usefulness ratings (95 percent rated these as somewhat or very useful), while the relationship of student achievement to variations in teacher qualifications got the lowest (71 percent rated somewhat or very useful). When asked for their rating of satisfaction with the manner in which NCES handled the state ranking or comparison issue, 40 percent of these respondents said they were very satisfied, another 40 percent said they were satisfied, while the remaining 20 percent said they were not satisfied or gave another response.

Respondents in the national educator and policymaker samples were also invited to comment on the reports and other materials prepared by NCES and to suggest ways in which the next cycle of NAEP reporting could be enhanced. The majority of respondents in both of these samples felt that NCES had done a good job. Some felt, however, that the data were still too voluminous and confusing and that greater efforts had to be made to simplify the reports. Others suggested that future cycles would be enhanced if NCES provided more analysis to show clearly how the NAEP data could be used to improve instruction.

The next cycles of data collection will afford another opportunity to see how these reports meet the continuing needs of analysts and policymakers. They will also permit an evaluation of the relative utility of the NAGB achievement levels as a reporting device.

How were the results of the TSA reported by the press?

Included in the materials distributed at the June 6, 1991 press conference that presented the results of the 1990 NAEP TSA were press releases designed to sound an alarm about mathematics achievement in America. Secretary of Education Lamar Alexander stated: "The big news is: none of the states are cutting it." Press releases by NAGB board members included several opinions that reinforced the same theme, beginning with a statement by Richard A. Boyd, chairman, pointing out that the results "do not portray a happy picture....let's use today's results to examine where our deficiencies may lie and then get to work to correct them." In another statement Francie Alexander, then a member of NAGB, said, "Unfortunately, the data show we have an enormous way to go even in the states which appear to be doing *relatively* better than others." Carl J. Moser, another NAGB member, focused on the way the factors of home and school interact and said, "What we need is a national effort to promote learning and to value it....It means a change in our values and expectations." The common theme was to motivate the American public to focus their attention and efforts on improving mathematics achievement.

This theme was picked up and reinforced by others in the education policy community. For example, CCSSO headlined its release with "National Mathematics Campaign Imperative," and quoted Gordon M. Ambach, CCSSO's executive director: "The results from the 1990 math NAEP indicate a tough, uphill journey for students...a national campaign is imperative."

By and large, the press reported accurately the central message portrayed by Secretary Alexander (and echoed by many others, including NAGB board members and spokespersons for many of the states)--that mathematics performance is inadequate and needs to be improved.

Typical headlines included "U.S. Students Weak in Math, Study Reveals: Most computer-age kids know only basic arithmetic" (*San Francisco Chronicle*), "Report charts U.S. students' shortcomings in math" (*Philadelphia Inquirer*), and "Math's a problem, tests show; 'we compare favorably,' but educators aren't happy" (*Rocky Mountain News*, Denver).

State comparisons. When it came to the state comparisons, however, four of the most influential newspapers in the nation, the *New York Times*, *Wall Street Journal*, *Washington Post*, and *USA Today*, rank-ordered the states in their news articles despite caveats from both NCES and Secretary Alexander not to do so. The *New York Times* warned against "using the data as a simple ranking because of the many variables involved in the calculations and the overlapping performance of many states," while the *Post* mentioned that many of the tabled differences were not statistically significant. Neither paper, however, provided any further guidance as to *which* of the differences were significant or not significant. Moreover, the *USA Today* article contained the incorrect interpretation that "Eighth graders in North Dakota, Montana, Iowa and Nebraska know 15 percent to 20 percent more math than those in Louisiana, North Carolina, and Washington, D.C." Although the mean scale score of top-scoring North Dakota's eighth graders *is* 21.6 percent higher than the mean scale score of the District of Columbia's eighth graders, for the *USA Today* interpretation to be correct would require that a NAEP scale score of zero represent total absence of knowledge and that the NAEP test measure the whole of subject area knowledge.

In a review of 55 articles published in the top 50 newspapers (by circulation) during the week the results were released, it was found that state rankings were mentioned in about two thirds of the articles. Ten accounts presented the average scores for all states in table form, either ordered by state rank or with rank indicated, and two others presented partial or complete lists of the states ranked by the percentage of students scoring at Level 300. Fewer than half of these articles--and only three of those that used tabular displays--provided appropriate interpretation of these statistics.

By comparison, when describing the ranking of a *single state*, at least half the reporters were able to formulate appropriate interpretations based on the language provided by NCES and their own state departments of education. For example, the *Hartford Courant* reported that "Allowing for possible statistical errors, researchers said that only five states scored significantly higher than Connecticut and 12 others had roughly equal scores," and the *St. Petersburg Times* wrote that "Nineteen states scored higher than Florida, sixteen made similar marks and four ... posted significantly worse scores." By contrast, however, The *Miami Herald Tribune* reported the Florida data as follows: "37 states, 2 territories and the District of Columbia participated and were ranked in 3 groups, with 19 in the first group, 17, including Florida, in the second, and 4 in the third." These "groups" were not part of the analysis or reporting design, as implied by the article, but simply the numbers of states who ranked higher, the same, or lower in relation to Florida.

Breadth of coverage. The 1990 NAEP produced enormous quantities of data, and even the highly condensed press release issued by NCES presented highlights of more than a dozen separate aspects of national and state performance. Most newspapers, however, published only a single article on the results. It is not surprising, therefore, that the news reports generally failed to capture the full richness of the material. Most of the articles focused on the overall results for the nation the paper's home state, or both, but only a few provided detailed discussions of performance differences for subpopulations within their state or between their state and others. Similarly, about half the articles touched on the effects of TV watching or the amount of time spent on homework, and about a quarter reported briefly on the positive effects of teacher specialization and parent educational levels. However, significant coverage of some of the most interesting and potentially useful aspects of these contextual data, such as the relationship of student performance to differences in mathematics curriculum or teacher preparation, was scant at best.

How do the respondents evaluate the press coverage received by the TSA?

When asked how the news media handled the release, 44 percent of the state assessment directors and mathematics specialists said that the press did a good or acceptable job, 12 percent rated the media's performance as fair or mixed, and 15 percent said that the media did a poor job. Another 26 percent said that the June 1991 data release was a media "non-event." Nearly two thirds of these respondents said that the news media accurately conveyed the messages their states wanted to deliver about the results, while one quarter said the messages were not reported accurately. Twenty-eight percent said they were surprised by the news media coverage, most often the lack of coverage in their state.

The majority of the national educators and policymakers sampled also agreed that the information presented by NCES and others was reported accurately by the press. Some, however, expressed concern that the data were not well understood and that the coverage was too simplistic and incomplete.

How widely is NAEP known? By July-August 1991, how many of the potential consumers of TSA data had seen the results?

The state assessment directors and mathematics specialists, as well as those individuals selected for the sample of national educators, were necessarily familiar with NAEP because they had all participated in the development and/or the implementation of the 1990 TSA. Respondents in the national policymaker sample also all proved to be familiar with the NAEP program. Only 65 percent of the state sample and 54 percent of the local sample, however, recognized NAEP by name when interviewed in April-May 1991--prior to the release of the TSA state-by-state results. About another quarter of each of these respondent groups thought that they had heard about a national program to measure student performance at the state level. Respondents at the local level particularly were unlikely to be aware whether their own state had participated in the 1990 TSA. (Forty-two percent of the local respondents from participating states thought that their state had participated, but so did 33 percent of the local respondents from non-participating states.)

As noted earlier, the local sample, which is to be redrawn at each data collection, is expected to provide the main indicator of whether or not NAEP became more widely known following the release of the TSA results. (Respondents in the other samples, which are being maintained across time, do not serve as well for this purpose because they become sensitized to NAEP through the interview process itself.) For logistical reasons, however, it was not feasible to draw and interview a new sample of local respondents during the summer data collection when many district offices are closed. Consequently, the question of whether NAEP has increased in visibility as a result of the publicity surrounding the release of the 1990 data will largely have to await the results of the third data collection cycle in January-February 1992.

The July-August interviews did show, however, that about 70 percent of the state sample respondents (as well as all of the national educators and policymakers, state testing directors, and state mathematics specialists), had seen or heard about the 1990 TSA results by the time of the summer interviews in July and August 1991. More than half of the state sample respondents had seen something in the newspapers, and similar numbers had seen the NCES executive summary, the large composite report, and (among participating states) the report prepared for their own individual state.

Among those state sample respondents who had seen the results, more than three quarters felt that the results had been presented in a way that allowed them to understand and interpret them appropriately. Only two percent said that the results were *not* presented in an intelligible manner; the remainder had no opinion. Similarly, over 60 percent of the national policymakers gave high marks to the reports produced by NCES, although a few complained that the reports were too bulky and not as easy to read as people had hoped.

How highly do different respondent groups value the NAEP TSA as a monitor of education programs? What do they see as NAEP's major strengths and weaknesses?

All respondent groups in both interview cycles were asked to rate the importance of being able to measure state performance against a common standard like the 1990 TSA. The results, which are shown in Table 6, indicate that the majority of each respondent group evaluates such state comparative data as at least somewhat important. There also appears to have been a shift towards greater valuing of these data over the period when the TSA results were first released.

Of course, in evaluating these or any other changes that occur between interview cycles, it is important to remember that other factors in the contemporary education arena may be influencing the respondents as much as or more than the 1990 TSA. In particular, the accelerating discussion of national standards and the possibility of a national examination system have both received wide media coverage. Somewhat ironically, it is just because the TSA is concordant with much of what is happening in state and national education policy circles that it may be extremely hard to isolate any impact of the TSA program *per se*.

Table 6. How important is it to be able to measure state performance against a common standard like the 1990 Trial State Assessment?

	Very important	Somewhat important	Not very important	Don't know
National educators:				
April-May (N=19)	32%	42%	26%	0%
July-Aug. (N=25)	52%	40%	8%	0%
National policymakers:				
April-May (N=19)	26%	47%	16%	11%
July-Aug. (N=18)	17%	67%	0%	17%
Assessment directors and math specialists:				
April-May (N=96)	39%	38%	24%	0%
July-Aug. (N=102)	39%	45%	16%	0%
State sample:				
April-May (N=54)	24%	56%	19%	2%
July-Aug. (N=74)	41%	41%	15%	4%
Local sample:				
April-May (N=110)	21%	48%	26%	5%

More than three quarters of the state assessment directors and mathematics specialists, and about half of the state sample respondents, however, felt that there was some other testing program or some other form of assessment that was more important for evaluating student performance in their own state. State tests, but also classroom assessment by teachers, and, from another perspective, performance-based assessments, were highlighted as more relevant forms of testing. Reasons given for these preferences included better alignment with local or state curriculum, greater diagnostic or instructional utility, more applicability to program assessment, and the ability to provide feedback at the individual, classroom, or district levels.

Specific problems cited as limiting the utility of the 1990 TSA included the fact that the test remains too traditional and does not sufficiently reflect the recommendations of the NCTM standards, a too small or improperly constructed sample, and the inability to make statements about performance at the local level. More than three quarters of the state assessment directors and mathematics specialists also indicated that they would like to see more state involvement in policy-setting for NAEP and the NAEP TSA. These respondents would like to see greater state input into test development and design, standards setting, and the reporting of results.

Do the respondents think that the NAEP state assessments should be continued after 1992? Would they like to see the NAEP program expanded to make results available for individual districts, schools, or students?

As part of the July-August 1991 interviews, the national policymakers sample (who represent those most intimately involved in NAEP's reauthorization) were asked whether they supported the continuation of the NAEP state assessments into 1994. Seventy-eight percent said that the program *should* be reauthorized, 22 percent said they hadn't yet formed an opinion (in some cases because they were awaiting the results of the evaluation of the 1990 TSA), and none said categorically that it should not be continued. Eighty-six percent, however, felt that the program should remain on a trial basis in 1994; the remaining 14 percent were equally divided between those who thought the program was ready to be given permanent status and those who had not formed an opinion.

The question of whether to expand the NAEP program to provide scores at the district, school, or individual level was put to the state and local respondents as part of the April-May 1991 interview cycle. As shown in Table 7, less than a third of the respondents in each of the assessment directors/or mathematics specialists, state, and local samples felt that it would be worth expanding the NAEP program to the district level if the costs of this expansion had to be borne by the states and the districts. Respondents in the local sample were least supportive, with only 15 percent endorsing such an expansion; however, there was also a high proportion of "don't know" responses among both the state and local samples.

Table 7. Assuming that the costs would have to be borne by the state or the districts, do you think it would be worthwhile expanding NAEP data collection in your state to generate data for individual school districts? Individual schools? Individual students?

For local respondents, the wording was altered to "Assuming that at least part of the costs would have to be borne by the local districts..."

	Assessment directors and mathematics specialists (N=96)	State sample (N=54)	Local sample (N=110)
Expand to district level?	26%	30%	15%
Expand to school level?	20%	24%	10%
Expand to individual level?	9%	20%	6%

Those who favored the expansion cited NAEP's unique ability to measure local performance on the same metric as state and national performance. Those who opposed expansion felt that the NAEP program would be too expensive and would only duplicate information that schools and districts already have from other assessment programs.

Have the 1990 TSA results given rise to meaningful discussions among educators and policymakers at the national, state, or local levels?

In July-August 1991, all respondents were asked whether the 1990 TSA results had provoked discussions among groups with which they were likely to be in contact. Thus the state assessment directors and mathematics specialists and the stat[illegible]ple respondents were all asked about discussions taking place within their own states, while the national educators and policymakers were asked about the national scene. As shown in Tables 8 and 9, it would appear that the results did provoke a reasonable amount of discussion, particularly in participating states and in the U.S. Congress. State-level discussions predictably tended to focus on the relative performance of the respondent's state compared to other states or on the relative performance of subgroups within the state. In a number of cases, however, the discussion appears to have gone beyond a simple consideration of the numbers and to contribute to proposals for educational reform.

Table 8. As far as you are aware, did the release of the TSA results provoke discussions about your state's math performance amoung any of the following groups?

	Assessment directors and mathematics specialists			State sample*		
	P (N=76)	NP (N=26)	Total (N=102)	P (N=42)	NP (N=8)	Total (N=50)
your state Board of Education	54%	19%	45%	50%	38%	48%
your governor's office or state legislature	28%	35%	29%	45%	0%	38%
your State Department of Education	84%	46%	75%	71%	38%	66%
local districts in your state	24%	8%	20%	24%	13%	22%
colleges and universities in your state	30%	8%	25%	21%	0%	18%
education organizations in your state	49%	15%	40%	57%	38%	54%
the general public in your state	20%	4%	16%	14%	0%	12%

Legend: P = 1990 TSA participant; NP = nonparticipant

* Asked of those who had seen or heard of the 1990 TSA results only.

Table 9. As far as you are aware, did the release of the TSA results provoke discussions about state or national math performance among any of the following?

	National educators (N=25)	National policymakers (N=18)
the President's office	40%	39%
the U.S. Congress	36%	83%
the U.S. Department of Education	40%	44%
state or local education agencies	36%	NA

What evidence is there that the TSA is beginning to influence state mathematics instruction or assessment?

The state assessment directors and mathematics specialists are probably the respondents most in a position to observe early influences of the TSA on state mathematics instruction or assessment. In fact, among the 87 percent of these respondents who indicated that their states were making or considering changes in their mathematics *instructional* programs as of July-August, 1991, 29 percent (including 35 percent of those from states that participated in the 1990 TSA and 13 percent of those from states that did not participate) indicated that the TSA results had influenced the change process, primarily by supporting changes already in process or by acting as a catalyst for change. Similarly, among the 85 percent of respondents whose states were making or considering changes in their mathematics *assessment* programs, 37 percent (including 43 percent of those from states that participated in the 1990 TSA and 21 percent of those from states that did not participate) credited the TSA with influencing these changes. In fact, respondents in 14 states indicated that their states were developing assessments that would use NAEP-like items or would parallel NAEP in other ways.

The other three respondent groups in the July-August interview cycle were asked in more general terms whether they anticipated that the 1990 NAEP state comparative data would result in any changes in mathematics instruction or assessment at the state level.[2] Forty-two percent of the state sample respondents, but only 28 percent of each of the two national sample groups thought that this was likely. All these groups, however, included a high percentage of "don't know" or "too early to tell" responses to this question.

Overall, was the 1990 NAEP TSA considered a useful exercise?

For a summary measure of the impact of the 1990 NAEP TSA, the state assessment directors and mathematics specialists and those respondents from the state sample who had some exposure to the TSA results were asked whether the overall impact on education in their states was positive or negative, while the respondents in the two national samples were asked to evaluate the success of the 1990 NAEP mathematics assessment in providing information that would help educators and policymakers to improve U.S. education. As shown in Table 10, as of July-August 1991 only a small proportion of the state-level respondents felt that the overall impact of the TSA on education in their state was negative, but the remainder (excluding those who had no opinion) were fairly well split between those who saw the impact as positive and those who saw it as mixed or very limited. It is noteworthy, however, that among the more knowledgeable assessment directors and mathematics specialists, respondents from states that participated in the 1990 TSA were much more likely than respondents from nonparticipating states to rate the overall effect as positive (54 percent compared to 15 percent).

---

[2] For the state sample respondents, the question was phrased in terms of anticipated changes in their own state.

Table 10. Overall, was the impact of the 1990 NAEP TSA on education in your state positive or negative?

	Assessment directors and mathematics specialists			State sample*		
	P (N=76)	NP (N=26)	Total (N=102)	P (N=42	NP (N=8)	Total (N=50)
Positive	54%	15%	44%	33%	38%	34%
Negative	3%	4%	3%	12%	0%	10%
Mixed or very limited	37%	65%	44%	36%	13%	32%
Don't know	7%	15%	9%	19%	50%	24%

Legend: P = 1990 TSA participant; NP = nonparticipant

* Asked of those who had seen or heard of the 1990 TSA results only.

Respondents in the two national samples, as shown in Table 11, were most likely to rate the 1990 NAEP mathematics assessment (national and state) as moderately successful. A sizable proportion of the national educators (28 percent), however, rated it as not very successful, citing their pessimism that it would have any impact on the way education is conducted in the states.

Table 11. From what you know at this point, how successful do you think the 1990 math assessment has been in providing information that will help educators and policymakers improve U.S. education?

	National educators (N=25)	National policymakers (N=18)
Very successful	8%	11%
Moderately successful	60%	72%
Not very successful	28%	6%
Don't know	4%	11%

Summary and Conclusions

Overall, it would appear that the survey respondents are cautiously optimistic about the potential for the TSA results to create positive changes in instructional and assessment practices.

Although many would have liked to have seen these results receive even greater attention from the public and the media, the evidence suggests that the results *were* widely disseminated and that they gave rise to meaningful discussions among a number of bodies concerned with education policy. Moreover, a substantial minority of those respondents most in a position to know (the state assessment directors and mathematics specialists) reported that, only a few weeks after the release, the results were already beginning to influence state-level changes in mathematics instruction and assessment, primarily by lending additional support to reforms that were contemplated or already underway. Future data collection cycles will show whether this influence continues, or whether it fades rapidly once the flurry of media attention is past. Given that subsequent NAEP state assessments continue to keep these matters before the public eye, however, it seems more likely that the influence will continue or grow rather than diminish.

While the reporting of the 1990 TSA generally seems to have gone very well, the authors, having considered the materials prepared by NCES and having compared the contents of NCES reports with the media coverage they generated, propose two recommendations for future reporting cycles. *First, in order to encourage the press to give more in-depth attention to the wide array of interesting and potentially useful TSA contextual data, such as the relationship of student performance to differences in mathematics curriculum or teacher preparation, NCES should consider spreading the release of the data over several months.* Under this plan, the first report on the state-by-state results could be followed by individual reports concentrating on the relationship of outcomes with such variables as course-taking patterns, instructional time devoted to particular subjects, teacher experience and preparation, and school characteristics. It would be important, however, to ensure that the initial report was sufficiently inclusive to provide a fair context for the comparisons between the states.

Second, NCES should continue to provide the media with specific examples of interpretations that are (or are not) supported by the results as well as continue efforts to construct new tables and graphs that effectively display the data. These newly developed reporting devices should be presented for comment to the press or media specialists in a series of focus groups, prior to the release of future reports. Because of the sensitivity of the state comparisons, the 1990 TSA gave rise to more concerted efforts than ever before to help the public and the press correctly interpret complex statistical concepts such as confidence intervals and statistical significance. Continued experimentation is needed, and NCES appears to be the agency best situated to undertake this effort. Certainly the survey results indicate that the states would welcome any additional materials that NCES can provide to help them get the message of the TSA results across to local press, policymakers, and educators.

General Issues in Reporting of NAEP Trial State Assessment Results

Richard M. Jaeger
University of North Carolina at Greensboro

October 1991

Acknowledgement

George Bohrnstedt, David Cohen, Robert Glaser, Robert Linn, and Albert Shanker provided excellent suggestions in response to an earlier draft of this paper. It is a pleasure to acknowledge their assistance. Of course, they bear no responsibility for remaining transgressions committed by the author.

Executive Summary

This paper addresses issues surrounding the reporting of results of student performances on the Trial State Assessments (TSAs) of the National Assessment of Educational Progress (NAEP). It concentrates on reports intended to inform the public about the of results the NAEP TSAs. As a basis for conclusions and recommendations, press briefing materials prepared by the National Center for Education Statistics, Educational Testing Service, and the National Assessment Governing Board for release of the 1990 NAEP TSA results were reviewed in detail, as was the Executive Summary of the NAEP composite report (Mullis et al. 1991a). Reports of 1990 NAEP TSA results that appeared in three national newspapers and two local newspapers the day after release of the results were also analyzed.

The following conclusions were reached: Despite the substantial and first-rate work of NCES and ETS professionals responsible for NAEP, reports intended to facilitate public communication of NAEP TSA results (the Mullis et al. 1991a *Executive Summary* and various press releases) did not realize their public reporting objectives. Newspaper reports of NAEP TSA results published the day after the NCES/NAGB press briefings were rife with misinterpretations and erroneous conclusions. In addition, it could be argued that NAEP TSA reports prepared for public consumption did not satisfy the requirements of several of the 1985 *Standards for Educational and Psychological Testing*. Reports intended to communicate NAEP TSA results to state-level decisionmakers were factual and well organized. Although it appears that their communicative value could be increased, the usefulness of these reports is an empirical question. This question will be answered, in part, through follow-up studies currently planned and being conducted by the NAE Panel on the NAEP Trial State Assessments and, perhaps, through the additional studies recommended in this paper.

Recommendations for improvement of the reporting of NAEP TSA results are as follows:

1. NCES and ETS should conduct a series of workshops for media personnel prior to the release of future NAEP results for the purpose of educating such personnel about defensible and indefensible

interpretations of NAEP findings, and to determine the types of data reports and displays that best satisfy the reporting needs of the media and their audiences. Additional studies of data reports that facilitate or inhibit the public's understanding of NAEP TSA results should also be conducted.

2. The utility of NAEP TSA reports intended for use by policymakers should be assessed through follow-up studies. In the interim, inclusion of data displays that portray strengths of relationships as well as anomalous cases should be considered for reports of the 1992 NAEP TSA.

3. Public reports of NAEP TSA results should contain interval estimates of states' mean performances rather than point estimates.

4. Overgeneralizations of NAEP TSA results by the press should be countered by including specific warnings and examples of interpretations that *are not* supported by results in public reports of NAEP TSA results. Examples of interpretations that are supported by results should be contrasted with examples of those that are not.

5. Public reports of NAEP TSA results should include simple graphs that emphasize the overlap and uniqueness in states' distributions of scores. Data displays and tables should call attention to the variability, as well as the central tendency, of states' score distributions.

6. Relationships between NAEP TSA results and characteristics of states and students should be summarized in simple graphical displays that portray strengths of relationships as well as anomalous cases.

7. Effective strategies for describing the meaning of states' NAEP TSA means that fall between NAEP anchor levels should be developed and disseminated.

8. Effective strategies for identifying and describing the *educational* significance of differences between states' NAEP TSA means (and

other distributional characteristics) should be developed and disseminated.

9. Public reports of NAEP TSA results should include richer descriptions of the meaning of student performance at the NAEP anchor levels and larger numbers of example test items associated with each anchor level.

Background

This paper addresses issues surrounding the reporting of results of student performance on the Trial State Assessments (TSAs) of the National Assessment of Educational Progress (NAEP). A bit of background on the Trial State Assessments, the study that gave rise to this paper, and the focus of the paper is provided before these reporting issues are joined.

The Trial State Assessment and its Origins

According to Ambach (1991), the genesis of the state-by-state component of the National Assessment of Educational Progress was a 1984 recommendation by the Council of Chief State School Officers that states be allowed to voluntarily participate in NAEP. Two states, New York and Wyoming, first participated in the NAEP reading comprehension assessment in that year. Haertel (1991) notes that the recommendation was endorsed in a National Academy of Education report entitled *The Nation's Report Card* (Alexander et al. 1987) Acting on the recommendation, the 100th Congress included a provision for two Trial State Assessments in Public Law 100-297, known as the Hawkins-Stafford Elementary and Secondary School Improvement Amendments. The first TSA has been completed. This Trial State Assessment, with data collected in February 1990 estimated the performances of about 2500 eighth grade public school students from each of 37 states, 2 territories, and the District of Colombia on a sample of 137 predominantly multiple-choice mathematics items. Since matrix sampling was used in data collection, not every student attempted every item. More specifically, the 137 test items were divided into seven approximately equivalent blocks, and each tested student attempted items in three of these blocks (Koffler 1991). The 1992 TSA will be broadened to include testing of fourth graders in reading and mathematics, and will repeat testing of eighth graders in mathematics. State participation in these TSAs is voluntary; as noted above, about three-fourths of the states chose to participate in the 1990 TSA. Reporting of state results was also voluntary. All participating states and territories chose to have their results reported on June 6, 1991, together with the national NAEP results.

The Role of the National Academy of Education Panel

In addition to mandating the TSAs, Public Law 100-297 required an independent evaluation of the feasibility of the TSA's and the validity of their findings. The National Academy of Education was contracted by the National Center for Education Statistics of the U. S. Department of Education to conduct this evaluation in October 1989. The Academy empaneled a broadly representative group of researchers and policymakers to conduct the evaluation, and engaged the services of the American Institutes for Research to provide technical and staff support (National Academy of Education 1991).

The Origins and Purpose of this Paper

To supplement its own investigations, the National Academy panel commissioned a series of studies and papers spanning a wide range of issues, including a review of the administration of the 1990 TSA, a study of the political context of state NAEP, and the content and curricular validity of the 1990 NAEP mathematics items. This paper, entitled "General Issues in Reporting," is one of three commissioned in the area of reporting of results of the TSAs. The intent of the Panel in commissioning the paper is embodied in the following statement: "The aim of this commissioned paper is to examine some of the 'macro' level concerns about reporting state NAEP results" (Robert Glaser--letter, November 21, 1991).

Current legislation authorizing the National Assessment of Educational Progress defines the purpose of NAEP as a means to "improve the effectiveness of our Nation's schools by making objective information about student performance in selected learning areas available to policymakers at the national, regional, state, and local levels" (NCES 1991,1). In addition, NAEP has been described as "The Nation's Report Card," (cf. Mullis et al. 1991a), connoting additional responsibility for informing the public about the status of student performance in the learning areas assessed by the program. This paper examines issues surrounding the reporting of NAEP Trial State Assessment results to these audiences. Reporting in ways that foster valid interpretations of NAEP results and minimize unwarranted and inaccurate interpretations is a central focus of the paper. Because the content and format of score reports designed to inform state and national-level policymakers have been negotiated through intensive discussions over a period of months (Owen 1991a), greater attention will be paid to reporting for the benefit of lay audiences than to reporting for the benefit of policymakers.

Organization of the Paper

Appropriate practices of those who design, develop, market, use, and interpret the results of educational and psychological tests are specified in the 1985 *Standards for Educational and Psychological Testing*, published by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. It is fitting, therefore, to begin the substantive portion of this paper with a section that reviews the requirements of the *Standards* concerning reporting of test scores.

The next section addresses issues surrounding the reporting of NAEP Trial State Assessment results to the public. This section begins with a set of propositions concerning problems that will likely surround public portrayals of NAEP TSA results, and it concludes with a series of elaborations on those problems that includes suggestions of score reporting choices and formats that might ameliorate the problems. Evidence from newspaper reports of the 1990 Trial State Assessment and national NAEP results is used to illustrate problems and motivate recommended approaches to solution. The response of the public press is reviewed in the context of materials made available by the Department of Education at its June 5 and 6, 1991 press briefings.

The paper concludes with a brief review of the categories of information on the 1990 NAEP TSA that were reported to national and state policymakers by NCES in its composite report and individual state reports. This section of the paper considers these reports in light of the requirements of the *Standards* and provides recommendations concerning ways the reports might be modified in light of the *Standards*.

Requirements of the 1985 *Standards for Educational and Psychological Testing*

Although the 1985 *Standards for Educational and Psychological Testing* contain 16 standards that pertain to score reporting, only 6 are obviously applicable to reporting results of the NAEP Trial State Assessments. These standards are as follows:

2.10 Standard errors of measurement should be reported at critical score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported for score levels at or near the cut scores.

3.16 The score report forms and instructional materials for a test, including computerized reports and materials, should facilitate appropriate interpretations.

4.1 Scales used for reporting scores and the rationale for choosing them should be described clearly in test publications to facilitate accurate interpretation of scores by both the test user and the test taker. A publication should specify how scaled scores are derived from raw scores.

4.2 When scaled scores are used, the scale chosen should be consistent with the intended purposes of the test and should be described in detail. When raw scores are used, they should be demonstrated to have the same desirable properties as do scaled scores.

6.8 When test results are released to the media, those responsible for releasing the results should provide information to help minimize the possibility of misinterpretation of the test results.

15.10 Those responsible for testing programs should provide appropriate interpretations when test score information is released to students, parents, legal representatives, teachers, or the media. The interpretations should describe in simple language the meaning of the scores, common interpretations of test scores, and the ways the scores will be used.

Discussion

The applicability of these standards to reporting results of NAEP Trial State Assessments is addressed and illustrated briefly in the discussion that follows. The *Standards* are then cited in the more extensive discussions and examples contained in the next section of this paper.

Standard 2.10. Although NAEP scores are reported on a continuum ranging from zero to 500, considerable attention is drawn to four score levels defined by 50-point intervals, ranging from 200 to 350. These score levels serve as implicit cut points for classification of students even though the levels are normatively based. Standard 2.10 therefore applies to current reports of NAEP TSA results and will be even more obviously applicable when NAEP results are reported in terms of the "basic," "proficient," and "advanced" achievement levels currently under development by the staff of the National Assessment Governing Board.

At present, in virtually all NAEP documents that contain data tables and figures displaying TSA results, estimated proportions of students whose achievement is at or above the four emphasized "anchor levels" of the NAEP scale are accompanied by standard errors of estimate (cf. Mullis et al. 1991a) thus satisfying the requirements of Standard 2.10. It is not clear, however, that these standard errors include errors of measurement in addition to random sampling errors. Since for aggregated statistics, errors of measurement will likely be greatly exceeded by random sampling errors resulting from sampling of examinees and items, adjustment of standard error estimates to include errors of measurement might be irrelevant. However, explanatory footnotes that accompany tables containing standard error estimates should be made explicit concerning the consideration and inclusion of measurement error estimates.

Standard 3.16. This standard embodies the major focus of this paper: how to develop score reports that facilitate appropriate interpretations and reduce inappropriate or invalid interpretations. Problems that are likely to arise when the media interpret NAEP TSA score reports are discussed at some length in the next section of this paper. This standard provides additional justification for considering these problems and for the potential solution strategies advanced.

Standard 4.1. Both national and state-by-state NAEP results are reported on a scale that is derived through a complex application of matrix sampling and item response theory. The scale incorporates the presumption that proficiency in a subject taught in schools (such as mathematics) is a unitary construct that can be estimated and reported in terms of a single cross-grade scale. When applied to NAEP Trial State Assessment reporting, Standard 4.1 requires that the rationale and assumptions underlying the NAEP scale be explained clearly to policymakers and the public. The complexity of the scale development process and the assumptions that undergird its construction and use will likely preclude the development of adequate explanations for the public. As noted later in this paper, the task of informing the public about what NAEP results mean, and what they do not mean, is enormously difficult. The tensions that result from psychometricians' desires to be technically correct and the limited abilities of the press and the public to understand the derivation and meaning of essential technical arguments, are virtually impossible to resolve. Perhaps with greater effort, reports on NAEP results can move closer to satisfaction of Standard 4.1, but it is unrealistic to assume that the public has the ability or the desire to understand how the NAEP scale was developed. As a compromise, reports on NAEP results intended for the public might include examples of conclusions that are warranted by the scale, as well as conclusions that are not. Through such examples, the public might come to understand the meaning of the NAEP scale, if not its derivation.

Standard 4.2 The NAEP Trial State Assessment is intended to accomplish several purposes, some of which might best be served by the unitary cross-grade scale currently used for reporting student performance, and others which might best be served by establishing separate, within-grade performance scales. In a subject area with clearly definable content categories, such as mathematics, it might be desirable to maximize between-state discrimination by establishing a separate NAEP scale for each content area or for each grade level. These considerations have been raised in the past by several NAEP technical review panels. Although they are beyond the scope of this paper, Standard 4.2 certainly provides a rationale for engaging in additional discussion of these issues.

Standard 6.8. The next section of this paper contains an extensive review of public reports of the results of the 1990 NAEP Trial State Assessment and the interpretation of those reports by selected elements of the public press. Standard 6.8 provides a clear warrant for such a review.

Standard 15.10. This standard has implications for clear description of the NAEP scale and the anchor levels used to highlight student performance. It requires descriptions of anchor levels that adequately and accurately convey the meaning of student performances in terms that the public and policymakers can interpret validly. Most obviously, released test items associated with each anchor level should, through their item characteristic functions, content, and structure, appropriately characterize student performance at the anchor level they purportedly represent.

One strategy for minimizing misinterpretation of test results might be public release of defensible interpretations or public release of discussion that contrasted defensible interpretations with the most likely or frequently observed indefensible interpretations in press reports resulting from the June 1991 NAEP TSA reports.

Reporting to the Public

As noted earlier, because the National Center for Education Statistics and Educational Testing Service have discussed and negotiated the content and format of NAEP Trial State Assessment reports with state-level policymakers over a period of months, and as a result of these discussions, have produced reports that conform to the expressed desires of state-level policymakers, this paper focuses principally on public reporting of the results of the NAEP Trial State Assessments. In one sense, materials available for public reporting of NAEP TSA results are consequences of the decisions that shaped the individual state reports and the composite report intended for policymakers. One might therefore consider public reporting materials to be secondary effects of decisions that gave rise to the reports designed for state-level policymakers. Although an elaborate set of materials was prepared for the June 5 and 6, 1991 press conference and briefing that described NAEP TSA results to the media, the content and format of those materials were essentially constrained by data displays prepared for the primary NAEP TSA reports. It is therefore useful to consider in some detail the information provided to the media by the National Center for Education Statistics and the National Assessment

Governing Board, and interpretations of that information, as they appeared in major newspapers throughout the nation.

This section begins with a set of propositions concerning the problems of public reporting of NAEP Trial State Assessment results and a list of suggested consequences of those problems. It ends with an elaborated discussion of the propositions and consequences that incorporates some potentially ameliorating approaches to reporting and display of NAEP TSA results.

Propositions and Consequences

Proposition 1: *The press will demand simplicity in reporting, regardless of the validity of the inferences likely to follow from simple reports.* Consequences: It is difficult to balance the media demands for simple reports (such as a ranked list of states or a ranked list of state means on grade 8 mathematics performance) under the guise of "the public's right to know" and what the scientific community knows about technically defensible reporting that might support valid inferences.

Proposition 2: *Apparent differences in student performance do not always reflect real differences.* Consequences: Unless prevented from doing so, the press will likely report all apparent differences in student performance (e.g., state mean grade 8 NAEP mathematics performances) regardless of their validity. Reporting policies and practices might be formulated so as to prevent or reduce such errors.

Proposition 3: *The press, the public, and those with policy-making authority will likely overgeneralize NAEP TSA results, even if warned not to do so.* Consequences: It probably is not possible to report results in ways that prevent their overgeneralization. However, it should be possible to say more than "don't do that" and to accompany results with examples of correct and incorrect interpretations.

Proposition 4: *Means do not tell the whole story about students' NAEP TSA performances.* Consequences: Although the average performance of tested students in a state will be of interest to the press and might be of interest to the public, the policy value of such statistics, in the absence of others, is limited. Reporting policies and practices should illuminate and portray distributions of performance.

Proposition 5: *Aggregated test results often mask as much as they describe.* Consequences: When warranted and possible, the performance distributions of significant subpopulations should be prominently featured in NAEP TSA score reports. States with similar distributions of grade 8 NAEP mathematics

BEST COPY AVAILABLE

performance might differ substantially in the performances of their constituent subpopulations.

Proposition 6: *When testing agencies do not provide interpretations of test results, the public and the press will create their own. Occasionally, but not frequently, these interpretations will be accurate.* Consequences: If NAEP TSA results pique the public interest, the reasons underlying between-state disparities in student performance will also be of interest. In the absence of interpretations and speculations concerning underlying causes and their likelihood, the public and the press will form their own, typically without a sound scientific basis. Reports of NAEP TSA results should be accompanied by results of analyses of relationships with potentially explanatory factors. Reports on such relationships should be couched in relational, rather than causal, terms.

Proposition 7: *Reports of NAEP TSA results should not depend solely on the normative properties of the NAEP scale.* Consequences: Although the principal interests of the public and the press are likely to be comparative and therefore rooted in the NAEP scale, education policy-makers, as well as the public, should be informed about the meaning of various scale levels in terms of associated test items. This is in keeping with Ralph Tyler's original design for NAEP (Linn 1990). The misinterpretations associated with past reporting of anchor items (Linn 1990) must be reduced or eliminated by providing additional information on examinees' actual performances on those items.

Elaborations

This list of propositions concerning problems arising from public reporting of NAEP Trial State Assessment results constitutes a set of assertions concerning the behavior of the press and the public in response to release of these results. These propositions were formulated prior to the June 5 and 6, 1991 press briefing and conference and the appearance of articles in the June 7th editions of newspapers. As these convictions are elaborated and their implications for reporting of NAEP TSA results are discussed, alternative reporting strategies that might ameliorate the problems inherent in the propositions will be noted, and newspaper reports on the 1990 NAEP TSA in eighth grade mathematics will be cited and compared with the press materials provided by the National Center for Education Statistics and the National Assessment Governing Board.

Proposition 1: *The press will demand simplicity in reporting, regardless of the validity of the inferences likely to follow from simple reports.*

On June 7, 1991, the front page of *USA Today* contained a story headlined "U. S. math skills ring 'alarm'." The story included and referenced a table containing mean NAEP scale scores for the District of Columbia, 37 states, and 2 territories that participated in the 1990 NAEP TSA. The article and the accompanying table attempted to feed the national hunger for scores, proclaiming triumph for midwest teams and doom for those from the southeast; it served no higher purpose. The errors implied by Propositions 2 through 6 were committed in this *USA Today* article: Although the mean NAEP TSA performances of some states' eighth graders differed reliably, many did not. The *USA Today* article ignored this issue. The article was couched in terms of relative educational quality among the states compared, generalizing from students' average performances on a limited sample of mathematics test items to their overall performances in mathematics and, by implication, to the general quality of schooling in the states compared. The article stated, "Eighth graders in North Dakota, Montana, Iowa, and Nebraska know 15 percent to 20 percent more math than those in Louisiana, North Carolina, and Washington, D. C. says the survey, the first to compare eighth grade skills by state." State means were compared, and other features of the distributions of scores in states were ignored. Louisiana was identified as the state with the lowest mean. That the lowest-performing tenth of Louisiana's students matched or outperformed the lowest-performing tenths of students in several other states was not mentioned. That North Carolina, the state prominently characterized as scoring second from the bottom, outscored only one state in the numbers and operations content area and the measurement content area but outscored three other states in the algebra and functions content area, was ignored as well.

Prominent policy-makers in Washington were consulted for comments and explanations. As stories in other major newspapers confirmed, excuses were abundant in low-ranked states. For example, as reported by the Greensboro, North Carolina *News and Record*, Suzanne Triplett, the state's Assistant Superintendent for Research and Evaluation, blamed the results on "a lack of parental involvement, too much television, and a poorly focused curriculum"; the state's Governor, Jim Martin, blamed North Carolina's poor performance on its bureaucratic school structure: "These test scores show what every other test shows: rigid, centrally controlled schools do not produce good results, no matter how much money we pour into it" [sic]. Some self-satisfaction was evident in states that ranked high, for example, in a story headed "Graded on a national curve, Hoosier math scores are OK," The *Indianapolis Star* quoted state Superintendent of Schools H. Dean Evans as saying, "As we moved up into a higher order of expectations of students, we do less well, but the nation also does less well and we do generally better than an awful lot of states." In fairness, it should also be noted that Mr. Evans reacted to Indiana students' performances by stating, "If you were to take solace in this achievement, it would be equivalent to being happy that we had been assigned to the upper-deck seats on the *Titanic*," indicating full realization that Indiana students did not perform well on the eighth grade mathematics assessment despite the state's adequate normative performance. Mr. Evans attributed Indiana's superior showing on the lowest anchor level of the NAEP Mathematics TSA to the state's recent emphasis on remediation and helping poorly performing students. A few of these explanations and excuses might be accurate, but the accurate explanations were probably lost in the fog created by the rest. If sufficient attention is afforded Propositions 2 through 6, this gloomy portrait of misrepresentation of NAEP TSA results in the press might be avoided on successive June 7ths, even if it was unavoidable in 1991.

Proposition 2: *Apparent differences in student performance do not always reflect real differences.* The press briefing materials authored by Emerson Elliott and Gary Phillips (1991) and the Executive Summary Report prepared by Ina Mullis et al. (1991a) went to substantial lengths in attempting to inform the media that simple ranking of state mean scores might well misrepresent the relative position of the states' NAEP TSA performances. It cannot be overemphasized that NCES and its prime contractor, Educational Testing Service, behaved responsibly by providing the press with information that was factual and informative on this point. In addition, they attempted to educate members of the press concerning the misleading consequences of publishing a simple ranking of states on mean NAEP mathematics scores. On page 12 of the Elliott-Phillips statement we find:

> ... simple averages can lead to incorrect conclusions about the achievement of one State in relation to another. The following comments will serve as a guide to the reader.
>
> First, all statistical surveys, including the National Assessment of Educational Progress, contain error due to sampling. This sampling or "standard error" of a survey estimate is a measure of precision with which an estimate from a particular sample approximates the population values of what is being measured. Survey results are frequently expressed in seemingly precise averages (e.g., average achievement of 219). In fact, however, there is a scientifically calculable variance above and below that

average -- a calculation that is unique for each value. The interval equal to twice the standard error above and an equal amount below the average is called the "95 percent confidence interval." This means that there is a 95 percent chance that the interval includes the real value of the population.

Thus, simple averages can be misleading. Suppose that the 95 percent confidence interval for State A overlaps the average for State B. Then there is not a statistically significant difference between the two averages; statistically, the two could be the same or even reversed in rank order.

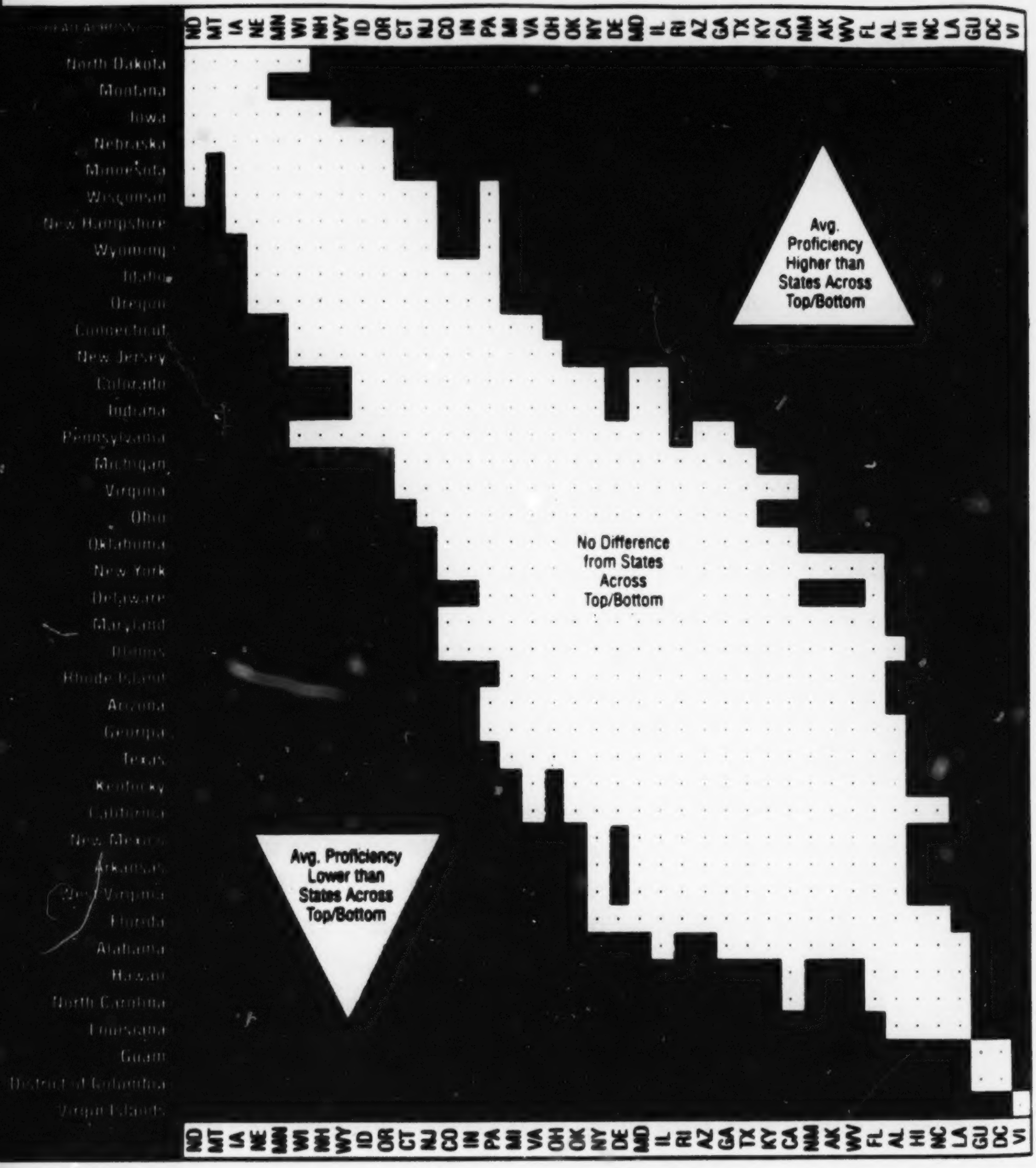
On page 15 of the Mullis et al. report, a graph (Figure 2) through which Educational Testing Service attempted to portray differences in state mean scores that were, and were not, statistically reliable was explained as follows:

> FIGURE 2 provides a method for making appropriate comparisons in average overall mathematics proficiency across the states (including the District of Columbia) and territories participating in NAEP's 1990 Trial State Assessment Program. The states are listed by overall average mathematics proficiency [NB: thus giving the press the information needed to construct a simple ranking of states]. However, the information presented in FIGURE 2, which uses appropriate tests of statistical significance to determine when average proficiency between states differs, shows that it would be quite misleading to assign numerical rankings (1 to 40) to these results. As can be seen, the pattern for most states is one of having lower average proficiency than some states, the same average proficiency as some states, and higher average proficiency than some states.

Unfortunately, most newspaper articles ignored this information. Figure 2 in the Mullis et al. (1991a) report (also called "Chart A" in the Elliott and Phillips press release materials), which came to be known as the "pantyhose chart," was judged by the press to be far too complex for public consumption. It is not clear that members of the press who attended the NCES/ETS press briefings understood the chart themselves, despite substantial attempts at explanation and clarification by NCES and ETS statisticians. Figure 2 is reproduced below.

FIGURE 2

Comparisons of Overall Mathematics Proficiency Based on Appropriate Tests of Statistical Significance



Note: Reading across, from left to right, this chart shows whether the average proficiency of each state or territory is lower than, the same as, or higher than that of other participants.

*Significance determined by an application of the Bonferroni procedure based on 780 comparisons by comparing the difference between the two means with four times the square root of the sum of the squared standard errors.

For any given state:

■ Overall average proficiency statistically significantly higher than comparison state.

□ No statistically significant difference from comparison state.

■ Overall average proficiency statistically significantly lower than comparison state.

BEST COPY AVAILABLE

As noted earlier, *USA Today* printed an alphabetically ordered table of state mean scores together with a list of state ranks, and totally ignored the issue of statistically unreliable differences between state means. In its narrative summary, the article in *USA Today* stated "On a scale to 500, North Dakota eighth graders scored 281, Montana 280, and Iowa 278. The Virgin Islands at 218, and D.C. and Guam, both at 231, are on the bottom." Not only did the article present these statistics without regard for indistinguishable differences between states, it did not identify the state "scores" as averages.

In an article titled "U. S. Youth Fail Math Test," appearing on the front page of the June 7th *Washington Post*, a staff writer reported that:

> Eighth graders in the District of Columbia ranked next to last among the 37 states and three territories that participated in the study--only the Virgin Islands did worse ... Virginia and Maryland students ranked in the middle of the survey. ... four rural states in the heartland--North Dakota, Montana, Iowa, and Nebraska--performed best in the National Assessment of Educational Progress, a congressionally mandated series of tests. ... The top state score, 281, went to eighth graders in North Dakota, although Montana, Iowa and Nebraska trailed by a statistically insignificant 1 to 5 points.

The article in the *Washington Post* included a table in which states were ranked according to their mean scores, with the state means and ranks indicated. However, the table carried the following footnote, indicating understanding on the part of the staff writer that states with very similar mean scores cannot be distinguished reliably: "NOTE: Experts explain that slight differences in test scores that affect ranking are not statistically significant." Although this footnote might signal readers that states with nearly identical means should not be regarded as having scored differently, readers would have no idea whether mean differences of one point or five signal trustworthy distinctions.

The *Indianapolis Star* article, cited earlier, did not provide information on states' comparative performances on the NAEP Trial State Assessment, beyond noting that "Hoosier eighth graders scored above average among the 36 states participating in the test" and "Just eight states scored better than Indiana, which was statistically tied with 12 other states. Fifteen states fared worse than Indiana." This report is essentially consistent with information provided in Chart A of the Elliott-Phillips (1991) press release. It appropriately acknowledges that the mean scores of some states are indistinguishable, either because they were identical or were close to being equal.

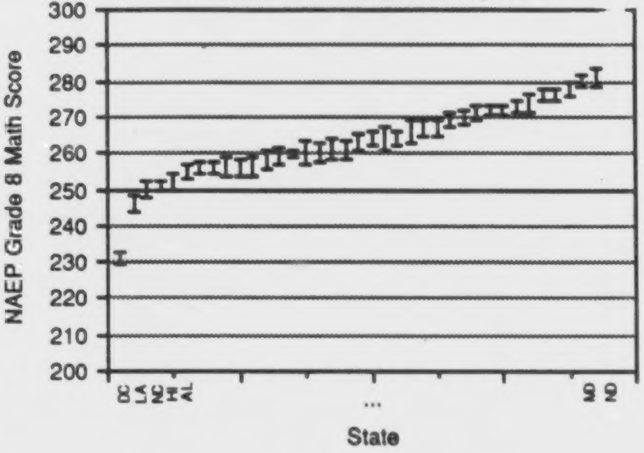
A front-page article in the June 7th *New York Times* titled "Eighth Graders' Math Survey Shows No State Is 'Cutting It' "included two tables containing state-by-state results. An abbreviated table on the front page reported that the eighth grade assessment had a "possible score of 350" and that the "national average was 261." This table listed ranks and means for the three states with the highest averages, and ranks and means for the two states plus the District of Columbia that had the lowest averages. In a related front-page narrative, the article stated, "Nine of the top 10 states, ranked by average score of their students, were in the Midwest or Northwest; New Hampshire ranked seventh. ... New York fell squarely in the middle, 20th, while Connecticut and New Jersey ranked 11th and 12th overall in the average score of their students." The article was continued on the back page of the fourth section of the *Times*. There the paper printed a full rank-ordered listing of states by mean score, together with state means and a graph showing the 5th percentile and the 95 percentile of scores in each state. Again the "best possible score" was stated to be 350. Although the back-page text of the *Times* article included the statement, attributed to Secretary Alexander that: "Listing these state-by-state figures 1 through 40 is, among other things, inaccurate," the paper proceeded to do just that. As was true of the *USA Today* article, no consideration was given to the statistical reliability of distinctions among states with similar mean scores.

The Greensboro, North Carolina *News and Record* printed a banner headline in its June 7th edition: "N. C. math skills don't add up." The related story contained a table titled "POOR IN MATH" that showed mean scores in reverse-rank order for the five lowest ranked states including the District of Columbia, and a "U. S. AVERAGE" score of 261. Once again, a newspaper reported state ranks without warning its readers about the issue of statistically unreliable distinctions.

If this sampling of five newspapers (including the three regarded as being the most influential in the nation) is representative, the attempt by Secretary Alexander to discourage simple ranking of states on the basis of their NAEP TSA means, and the efforts of those in the National Center for Education Statistics and Educational Testing Service, to convince the press that some apparent distinctions among state means should be ignored, were largely futile. Only the *Washington Post* included an appropriate warning that some tabled between-state differences were meaningless. Clearly, an alternative strategy for dealing with this problem is needed.

The error committed by the majority of papers reviewed could be avoided by releasing statistics on comparative state performances *solely* in terms of interval estimates, rather than point estimates of means and proportions. The press could be told that state means are not being reported because states with overlapping interval estimates exhibited mean student performances that are indistinguishable. The states could be listed in order of increasing interval upper bounds. Results could also be released in a graph such as the following:

Fig. 1 95 Percent Confidence Intervals on Mean Grade 8 NAEP Math Scores, by State



This graph clearly indicates that the average performance of public school eighth graders in the District of Columbia is reliably lower than that in all the states, but the average performances of eighth graders in Louisiana and North Carolina are indistinguishable, as are the average performances of eighth graders in North Carolina, Hawaii, Alabama, and Florida. The graph could carry this clear legend:

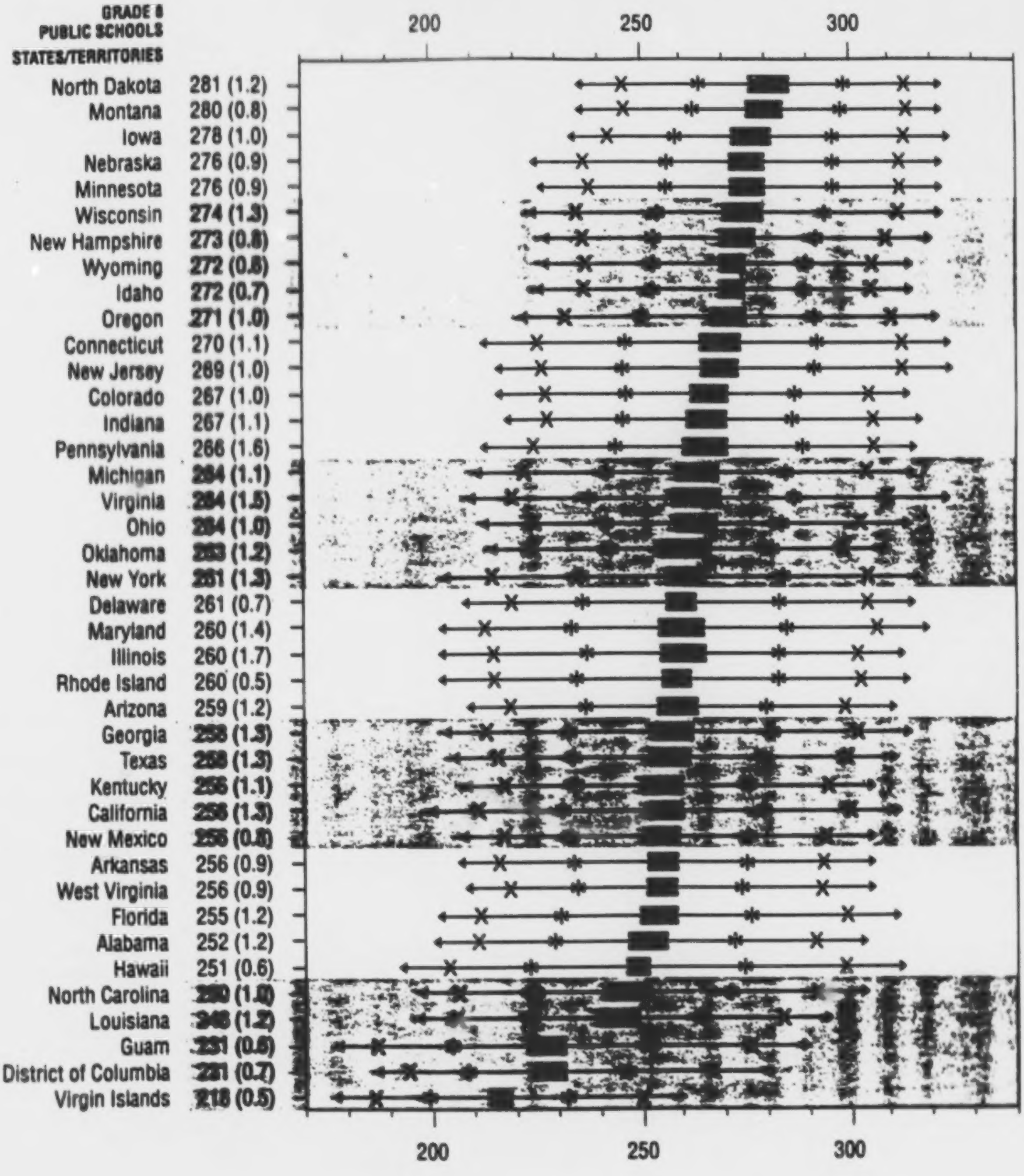
> The average performances of eighth graders in states with overlapping intervals should be considered equal. The average for one state is higher than the average for another *only* if the interval for one state is *totally* above the interval for the other.

Only reliable differences in average performance, such as those of Louisiana eighth graders and Alabama eighth graders, would be regarded as real differences. Note that the data used to create this graph were taken from Chart B of the Elliott-Phillips press release dated June 6, 1991 (Chart B was also shown on page 239 of the Mullis et al. (1991b) report, where it was labeled Figure 11.2; the figure from the Mullis et al. report is reproduced below because it provided clearer copy).

In their press releases, the National Center for Education Statistics and the National Assessment Governing Board provided several graphs (Charts A and B) that illustrated the point made by the graph shown above. However, both graphs provided additional information and were considered "too complicated" (Owen, 1991) by members of the press, thus confirming Proposition 1. Chart B incorporated an unfortunate error in placement of confidence bars around the mean scores of students in several states; the bands were placed asymmetrically, sometimes suggesting reliable between-state mean differences when none existed and at other times suggesting the absence of reliable mean differences when they were present. The graph proposed here has the advantage of simplicity and accurately conveys the idea that some differences between state means should be ignored. In this sense, it is consistent with the provisions of Standards 3.16 and 6.8.

FIGURE 11.2

Distribution of Overall Mathematics Proficiency
Organized by Average Proficiency



The shaded box indicates a simultaneous confidence interval around the average overall mathematics proficiency for the state based on the Bonferroni procedure for 780 comparisons (the average ±2.83 times its standard error). The remaining symbols indicate values of the 5th, 10th, 25th, 75th, 90th, and 95th percentiles of the distribution of overall mathematics proficiency.

BEST COPY AVAILABLE

If the news articles cited earlier are generalizable, some members of the press prefer tables to graphs for presenting statistical information to their readers. However, this conclusion is confounded by the absence of simple graphical summaries in the press materials distributed on June 5 and 6, 1991. A table that provided the same information as the graph shown above might be organized as follows:

State	The state mean score is between these values*	Number of states‡ with lower mean scores than this state	Number of states‡ with higher mean scores than this state
North Dakota	279 and 283	34	0
Montana	278 and 282	36	0
Iowa	276 and 280	33	0
Nebraska	274 and 278	30	0
.	.	.	.
.	.	.	.
.	.	.	.
Louisiana	244 and 248	3	33
Guam	230 and 232	1	37
District of Columbia	230 and 232	1	37
Virgin Islands	217 and 219	0	39

* If every eighth grader in a state had been tested, we can be 95 percent confident that the mean score for the entire state would fall within its interval.

‡ Includes two territories and the District of Columbia; students in 13 states did not participate in the Trial State Assessment.

The position advocated here might be regarded as extreme, since point estimates of state means would be withheld from the press and the public. However, if members of the press were given interval estimates of state means rather than point estimates, they could not misinform their readers by suggesting that states with statistically indistinguishable mean scores had different means. Information provided to the public would therefore be more accurate and less likely to be misinterpreted.

Since virtually all social and economic indicators are reported as point estimates (e.g., the gross national product, various crime rates, infant mortality rates, etc.), one might speculate that members of the press would regard NAEP results as uncharacteristically "sloppy," were they to be reported in terms of interval estimates. At worst, state-by-state

NAEP reporting might be considered by the press to be methodologically deficient, since it was "incapable" of estimating the exact average performance of students in each state. Whether this possibility presents a serious problem could be determined empirically and, if necessary, the criticism could be dispelled during a press briefing. Since nearly all responsible statistical agencies compute estimates of the variability of their indicators, such estimates could be used to compute confidence intervals around the kinds of well known statistics cited above. Members of the press could be taught that NAEP was being candid about its imprecision, whereas most government agencies hide data on precision of estimation in the final chapters of their lengthy reports. Comparisons of confidence intervals around NAEP estimates with those around well known social and economic indicators would probably show that NAEP results are no less precise than are the other indicators.

An additional problem associated with the recommendation to release state-by-state NAEP results in terms of interval estimates (rather than point estimates) is the difficulty of the concept of a confidence interval. The term "confidence interval" is statistical jargon that carries precise meaning for professional statisticians and no meaning for most members of the press or the public. The Gallup Poll and other professional polling organizations treat this problem by reporting point estimates, often accompanied by a statement such as "the margin of error in this survey is plus or minus 3 percent." This solution is imprecise, and probably misleading. The "3 percent" is a standard error of estimate; adding and subtracting this value from a reported point estimate provides an approximate *68 percent* confidence interval. Upon reading the typical Gallup Poll statement, most members of the public undoubtedly believe that they can be certain that the error due to sampling is no more than the reported "margin of error."

One of the recommendations provided in a latter section of this paper is to conduct research on what the public understands about NAEP TSA results and how the public interprets those results. Public interpretation of 95 percent confidence intervals, and the comparative efficacy of alternative labels and descriptions of confidence intervals, should be placed on this research agenda.

Proposition 3: *The press, the public, and those with policy-making authority will likely overgeneralize NAEP TSA results, even if warned not to do so.* The errors implied by this proposition take many forms. Perhaps the most egregious is the assumption that the rank order of states' NAEP grade 8 TSA mathematics means reflects the relative quality of schooling in those states. This inference can be found throughout the rhetoric that gave rise to the 1990 TSA and is unlikely to abate now that results have been released. For an example, see Phillips (1991), in which the Acting Associate Commissioner of the Education Assessment Division at the U. S. Department of Education's National Center for Education Statistics, after noting that "the NAEP Trial State Assessment is not a comprehensive monitoring system of the nation's elementary and secondary school system" went on to state (18):

> *Even though limited in scope, the Trial State Assessment will provide national and state policymakers, researchers, and the public with their first real glimpse of how well our state education systems are doing. ... State-by-state comparisons are the obvious first steps in developing a national information system that will shed some light on best practice and some heat on what doesn't work.* (emphasis added)

If the professional responsible for the nation's program of educational assessment assumes that NAEP TSA results can be used to infer the comparative quality of schooling in the participating states, imagine the overgeneralizations that will emerge from the lay press and public.

The press reports that followed release of the first TSA results focused more on the national NAEP results than on state-by-state comparisons. However, erroneous interpretations of national NAEP results very likely apply to TSA results as well, at least in the minds of newspaper readers. Examples abound. The June 7th article in *USA Today* includes the following overgeneralizations and misinterpretations: "Most students can't meet on-the-job demands for problem solving," "Only 5 percent [of high school seniors] are ready for college math," "Eighth graders in North Dakota, Montana, Iowa and Nebraska know 15 percent to 20 percent more math than those in Louisiana, North Carolina, and Washington, D.C.," "Boys score higher across the board, with the largest gap in the 12th grade (298-293)," and "More states are expected in the second round in February, when the scoreboard will be expanded to fourth grade." All of these statements are either untrue, or cannot be supported by the results of the 1990 National Assessment of Educational Progress. Consider them in order: There is no evidence that most jobs in the civilian labor market impose demands for mathematics problem solving beyond levels 200 or 250 on the NAEP scale. No information in the "mother NAEP" or NAEP TSA results supports the claim that "Most students can't meet on-the-job demands for problem solving." Similarly, the kinds of predictive validity studies needed to support the claim that "Only 5 percent [of high school seniors] are ready for college math" have not been conducted. Furthermore, this statement is inconsistent with a host of statistics on student participation in high-school mathematics courses and national statistics on college enrollment and attainment. According to the 1991 *Condition of Education* (U.S. Department of Education 1991a, 54), 6.2 percent of U.S. high school students completed a calculus course and another 12.8 percent completed an analysis or pre-calculus course. The 1988 *Digest of Educational Statistics* (U.S. Department of Education 1988, 248) indicates that 18.2 percent of students who were high school seniors in 1980 had attained a bachelor's degree by 1986. The 1990 *Condition of Education* (U.S. Department of Education 1990, Vol. 2, 36) reports that 58.8 percent of full-time, first-time freshmen in 1983 had attained bachelor's degrees by 1988, and that more than half the nation's high school graduates in 1987 enrolled immediately in college the next fall (U.S. Department of Education 1990, Vol. 2, 16). Finally, a May 1991 report issued by the National Center for Education Statistics (U. S. Department of Education 1991c, 26) indicated that only 21 percent of college and university freshmen were forced to enroll in remedial mathematics courses during the 1989-90 academic year. Since 79 percent of entering freshmen are not required to enroll in remedial mathematics courses, and over 60 percent of high school seniors enroll immediately in college the following year, the best evidence available indicates that almost 50 percent of high school seniors are "ready" for regular college-level mathematics courses, not the 5 percent claimed by *USA Today*.

The claim that "Eighth graders in North Dakota, Montana, Iowa and Nebraska know 15 percent to 20 percent more math than those in Louisiana, North Carolina, and Washington, D.C." is based on a faulty interpretation of the meaning of the NAEP scale. Although the mean scale score of North Dakota's eighth graders was 21.6 percent higher than the mean scale score of the District of Columbia's eighth graders, these scores provide no warrant to draw a conclusion about corresponding amounts of mathematics knowledge. To do so would require that a NAEP mathematics scale score of zero represent a total absence of mathematics knowledge, and the assumption that the NAEP mathematics test measured the whole of mathematics knowledge. Both assumptions are patently absurd.

In contrast to the newspaper's statement on gender differences, the Mullis et al. (1991a) Executive Summary contained the statement (12): "At grades 4 and 8, there seemed to be few gender differences, except males had higher average proficiency in measurement and estimation. However, at grade 12, males showed an advantage in every content area except algebra and functions. The gender differences in overall performance were most noticeable at the higher anchor levels on the scale."

Although the NAEP national results were sobering, they do not sustain the lead sentence in the *Washington Post* article of June 7: "A new national study of mathematics skills shows that most students have not mastered more than basic arithmetic ..." nor do they support the overgeneralization that "only 5 percent of [high school seniors] are prepared for college math." In fact, 46 percent of grade 12 students exceeded Level 300 on the NAEP scale and 5 percent exceeded Level 350. The Mullis et al. (1991a) summary interpreted the latter finding as follows (7): "... only 5 percent [of twelfth graders] showed an understanding of geometry and algebra that suggested preparedness for the study of relatively advanced mathematics."

Strategies for combatting this form of overgeneralization are not obvious. Perhaps reports of analyses that provide counterexamples to the simplest erroneous interpretations might be helpful. For example, the tendency to generalize from differences among average state scores on the mathematics assessment to differences in the quality of states' mathematics instruction might be reduced if it could be shown that states differed in terms of the opportunities of their students' to learn the mathematics content assessed in the 1990 TSA, and that there were demonstrable relationships between indices of instructional emphasis and students' average proficiencies. Unfortunately, the opportunity-to-learn data collected in the NAEP TSA provide a very weak indicator of the construct. A finding that between-state relationships with average student proficiency were strong would therefore be surprising. Nonetheless, a series of scatter diagrams that plotted, by state, the percent of students whose teachers reported "heavy emphasis" on a content area against the average proficiency of the states' eighth graders in that content area might be informative. Any non-zero relationship would suggest that states' students performed better in content areas their teachers chose to emphasize. Examples of such plots follow.

Fig. 2 Relationship Between Average Proficiency and Teachers' Reported Emphasis on Numbers and Operations, Across States

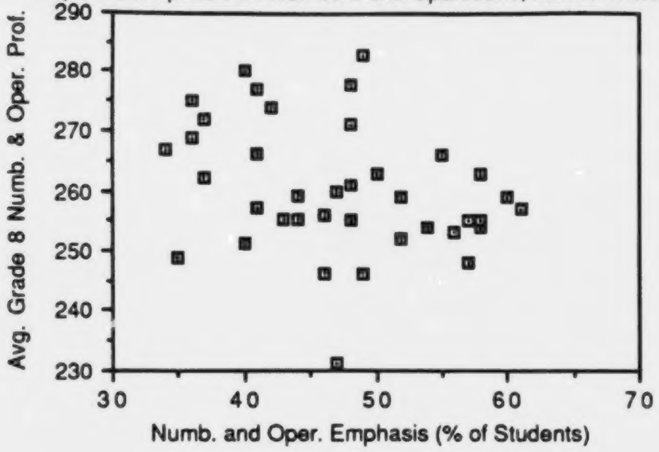
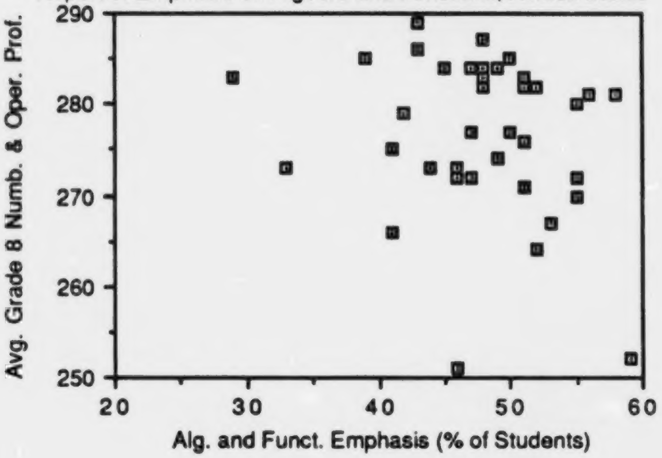


Fig. 3 Relationship Between Average Proficiency and Teachers' Report of Emphasis on Algebra and Functions, Across States



In these examples, the opportunity-to-learn indices predict very little of the variation in students' state mean grade 8 NAEP content area proficiencies. However, the graphs illustrate substantial variation in average instructional emphasis across states, suggesting that interpretation of NAEP TSA means in terms of the relative quality of the states' instruction in mathematics, rather than variations in states' content emphases in mathematics, might not be warranted.

It is unfortunate that the opportunity-to-learn data collected in the NAEP TSA are so limited. Ancillary data, possibly derived from statistics on state textbook adoptions, are unlikely to fill this void. Past studies of the content of delivered instruction (in contrast to the putative curriculum) have found large between-classroom variance components even when instructional materials, such as texts, were held constant (Schmidt et al. 1983). The current NAEP sampling design does not permit assessment of intact classrooms, and therefore precludes effective investigation of the true relationship between students' opportunities to learn and their NAEP performances (Koretz 1991).

Proposition 4: *Means don't tell the whole story about students' NAEP TSA performances.*

It is not uncommon for comparative statistical reports to concentrate on measures of the central tendencies of distributions of scores among the entities compared. However, exclusive reliance on means or other measures of central tendency often leads to unwarranted overgeneralizations concerning the comparative performances of the entities described and ignores important information embodied in other distributional statistics.

The press release materials for the NAEP TSA prepared by the National Center for Education Statistics and the National Assessment Governing Board clearly emphasized mean performances in the participating states. The same is true of the Mullis et al. Executive Summary (1991a). However, these sources also provided substantial information on other characteristics of the distributions of eighth graders' performances in the participating states. In particular, Charts B and C in the appendix to the Elliott and Phillips (1991) press release provided information on the states' NAEP TSA performances at seven major percentiles in both graphical and tabular form. The importance of these statistics in comparing states' performances can be gleaned from the following matrix of correlations among ranks of states' comparative performances at their respective means, medians, and 5th, 10th, 25th, 75th, 90th, and 95th percentiles, computed using data provided in Chart C of the Elliott and Phillips (1991) press release.

Correlations among the mean and selected quantiles of states' distribution of grade 8 NAEP mathematics performances

Correlation Matrix for Variables: X_1 ... X_8

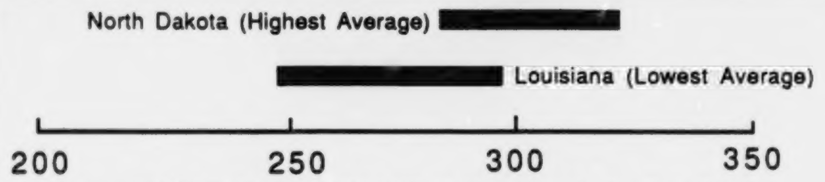
	RnkMEAN	Rnk5TH	Rnk10TH	Rnk25TH	Rnk50TH	Rnk75TH	Rnk90TH	Rnk95TH
RnkMEAN	1							
Rnk5TH	.938	1						
Rnk10TH	.949	.992	1					
Rnk25TH	.979	.967	.977	1				
Rnk50TH	.992	.924	.932	.974	1			
Rnk75TH	.975	.868	.878	.928	.976	1		
Rnk90TH	.92	.783	.793	.849	.913	.973	1	
Rnk95TH	.832	.673	.681	.742	.818	.908	.966	1

It is clear from this matrix that the rank order of states' eighth grade NAEP TSA performances varied, depending on the statistic used to describe those performances. Differences in rank order were particularly large when the ranks of extreme percentiles were correlated. However, some states' ranks changed appreciably even when less extreme percentiles (rather than their means) were used to describe their comparative performances.

The ranks of 22 of 38 states (58 percent; including the District of Columbia) changed by at least one when states were ranked using the 25th percentiles of their eighth graders' score distributions instead of their means. The ranks of 16 states (42 percent) changed by more than one, and the ranks of eight states (21 percent) changed by at least three. The ranks of 27 states (71 percent) changed by at least one when states were ranked using the 75th percentiles of their eighth graders' score distributions instead of their means. The ranks of 24 states (63 percent) changed by more than one, and the ranks of nine states (24 percent) changed by at least three, using the 75th percentile instead of the mean to rank them.

None of the materials prepared for distribution to the press at the June 5 and 6, 1991 briefings emphasized these points. Since considerable computation was necessary to determine that the states did not line up in the same way when compared using different statistics concerning their eighth grade students' math proficiencies, it is not surprising that the national press (as well as the local papers) restricted their-state-comparative remarks to mean performances. As noted elsewhere, only the *New York Times* included a graph that illustrated a few characteristics of the distributions of students' scores within states.

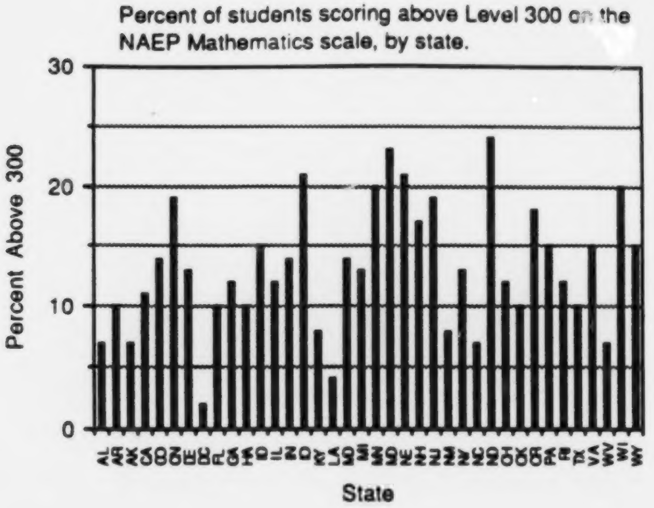
Distributional statistics would have been particularly useful in quantifying the Elliott and Phillips (1991) contention that "Differences among the states were very large," (p.8). Rather than citing the 35-point difference in NAEP scale means between North Dakota and Louisiana, they could have noted that only 15 percent of eight graders in the state with the lowest mean (Louisiana) outscored the average eighth grader in the state with the highest mean (North Dakota). The press might have found this information to be more compelling had it been presented in a graph such as the following:



Ranges of Eighth-Grade NAEP Math Performance in the Two States with the Lowest and Highest Sample Average Scores

The stability of state comparisons across alternative statistics is being examined more thoroughly in a study commissioned by the NAE Panel. In their report on the study they have underway, Robert Linn, Lorrie Shepard, and Elizabeth Hartka will further illuminate this issue. Pending the Linn and Shepard results, one can conclude that states' comparative NAEP TSA performances are not totally stable across statistics selected for reporting. This conclusion supports a recommendation that state comparisons be illustrated using several alternative statistics.

One alternative that provides some indication of the distribution of performance in each state was contained in the Mullis et al. (1991a) Executive Summary. Table 3 of that report, on page 18, contains percentages of students in each state with NAEP mathematics scores equal to or exceeding each of the four anchor points. Perhaps because of its complexity or its tabular form, the state-comparative aspects of Table 3 were ignored by the press. Since Level 300 is the anchor level most closely associated with the expected performance of eighth graders, a data summary that focused on this level might have been of greater interest than the more-complete Table 3. The following graphical portrayal of the percentage of students in each state who satisfied this expectation is one possibility.

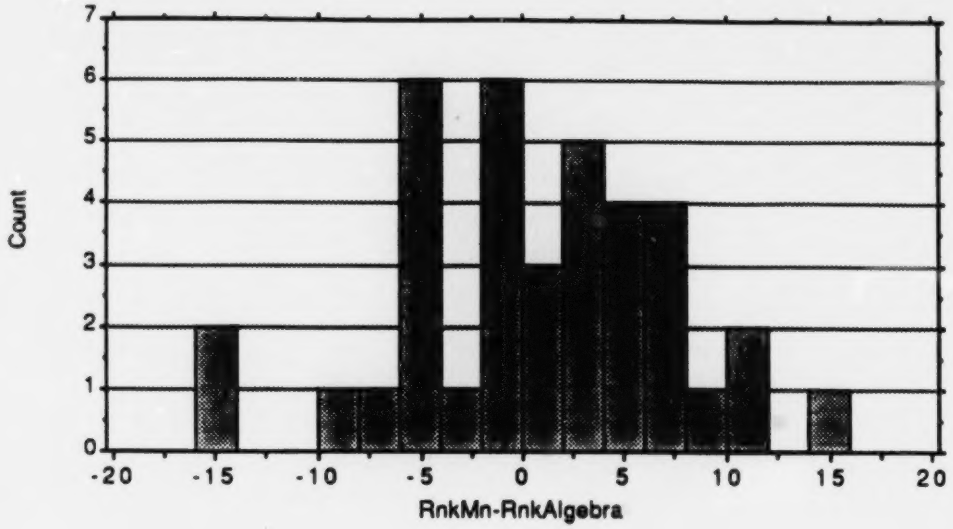


Were this graph to be used by the press, readers would have to be warned that the percentages were based on samples of students, and states with percents of eighth graders scoring above 300 that were within four percentage points of each other should not be regarded as having scored differently.

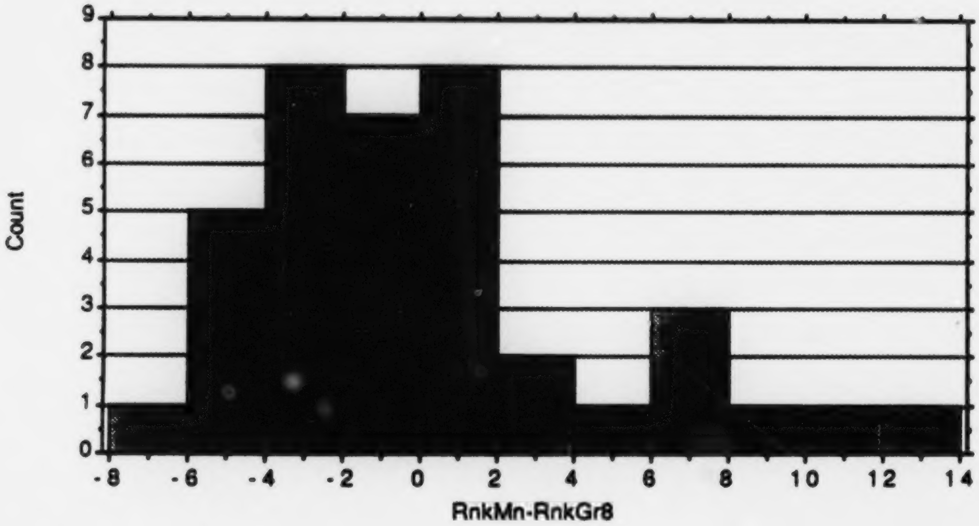
Proposition 5: *Aggregated test results often mask as much as they describe.*

Although there is a practical limit to the amount of information that the press can report and the information that the public will tolerate or attempt to assimilate, representation of states' comparative mathematics performances solely in terms of their entire populations of eighth graders masks information on the performances of their constituent subpopulations. This is particularly true in light of the unfortunate emphasis afforded states' comparative ranks by members of the press. Perhaps reporters could be convinced that states' overall rankings do not adequately convey their comparative effectiveness in providing mathematics instruction, and the public could be similarly informed, through data displays that illustrated differences between states' ranks based on their overall populations and their ranks based on important subpopulations. The following illustrative data displays show that states' comparative performances differ substantially when their ranks are computed from the mean performances of students enrolled in specific mathematics courses, instead of their overall populations of eighth graders:

The difference between states' ranks based on the mean grade 8 NAEP mathematics performance of all sampled students and the mean grade 8 NAEP mathematics performance of students enrolled in algebra



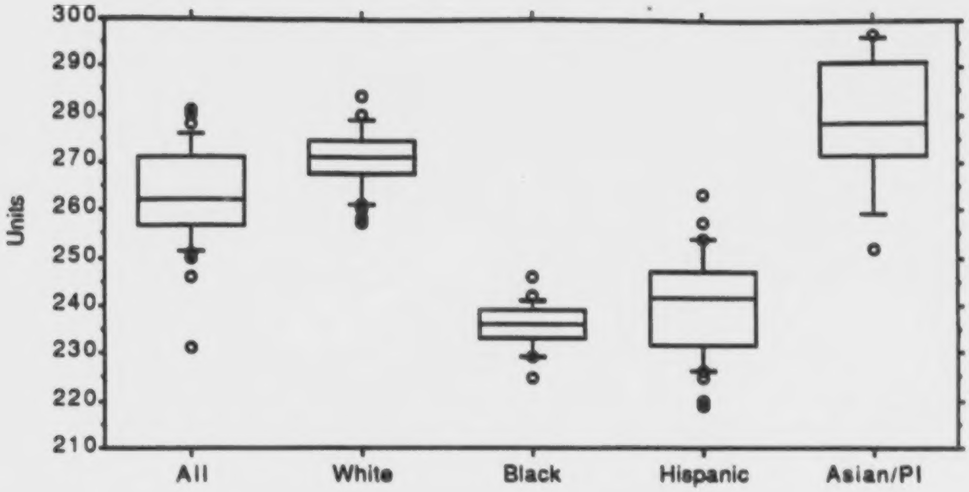
The difference between states' ranks based on the mean grade 8 NAEP mathematics performance of all sampled students and the mean grade 8 NAEP mathematics performance of students enrolled in 8th grade mathematics



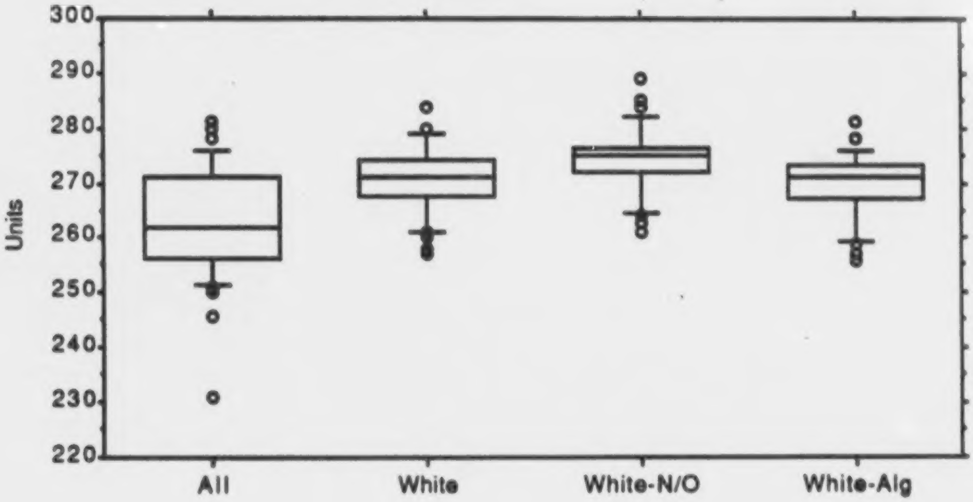
The first graph shows that states' ranks differ by as much as 15 (either higher or lower) when they are compared in terms of the average performances of their eighth graders enrolled in algebra, instead of their entire populations of eighth graders. Over 70 percent of the states' ranks changed by more than two when they were compared in terms of the algebra course subpopulation. The second graph illustrates changes in ranks that are almost as large when states were compared in terms of the average performances of their eighth graders enrolled in "eighth grade math," instead of their entire populations of eighth graders. These data displays support the contention that states' success in mathematics education depends in part on factors that guide their eighth graders to different course options, and on the effectiveness of those courses (Mullis et al. 1991a Executive Summary). These were the only statistics contained in summary documents or press releases that supported the construction of data displays for subpopulations of students in each state.

Had other data contained in the composite report (Mullis et al. 1991b) been summarized in press releases in the form of data displays, the overgeneralizations of NAEP TSA results by the press, cited earlier, might have occurred less frequently; e.g., a graphical summary that demonstrated the dependence of states' rankings on students' course enrollments might have prevented *USA Today* from concluding that "Eighth graders in North Dakota, Montana, Iowa and Nebraska know 15 percent to 20 percent more math than those in Louisiana, North Carolina, and Washington, DC." In addition, interesting relationships could have been illustrated. Two additional displays, based on data found in Tables 12.1 and 12.3 (266, 271, and 275) of the composite report illustrate this point.

Distributions of states' mean grade 8 NAEP mathematics performances, by racial-ethnic group of students



Distributions of states' mean grade 8 NAEP mathematics performances of all sampled students, white students overall and white students' mean scores on Numbers-Operations and Algebra

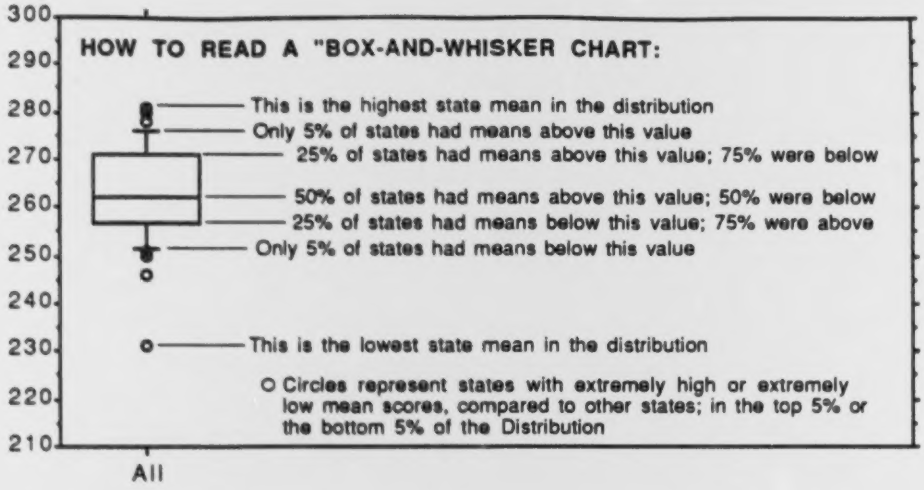


The boxplots in the first figure show that distributions of mean scores across states differed substantially for students of different racial and ethnic groups. Although materials prepared for the press reported this finding in narrative form, the distributions illustrate the magnitude of between-subpopulation differences far more clearly than any narrative report. They also show that distributions of state means for black students and Hispanic students overlapped substantially, as did distributions for white and Asian/Pacific Islander students. Moreover, they show that for the largest racial groups (whites and blacks) states were far more homogeneous in their average performances than was the case for overall samples of students. These findings support the interpretation that a substantial portion of state-by-state differences in mean NAEP TSA performances are attributable to differences in the racial-ethnic composition of their student bodies. Finally, if an average performance of 300 is regarded as an interim goal for eighth graders, these results show that states are far closer to achieving the goal for their white students and students of Asian/Pacific Island background than they are for their black and Hispanic students.

The second data display reinforces the contention that states' means differ far less for students in a particular racial-ethnic group than for students overall and that variation among state means is even smaller (although trivially so) when students' mean performances in a single mathematics content area are considered. White students' means in the numbers and operations content area or the algebra and functions content area, differ so little for the middle 50 percent of the states that they are, for practical purposes, indistinguishable.

These data displays have been prepared to illustrate the advantages of providing results of analyses for subpopulations of interest in ways that might readily communicate to members of the press. Whether these data displays are too complex for the public is an arguable question. To be sure, members of the press would have to be instructed in their interpretation. A description such as the following might be used:

> Box-and-whisker charts illustrate the shape of a distribution of scores, and indicate relationships among statistics that summarize a distribution of scores.
>
> A box-and-whisker chart should be interpreted as follows:


HOW TO READ A "BOX-AND-WHISKER CHART:
This is the highest state mean in the distribution
Only 5% of states had means above this value
25% of states had means above this value; 75% were below
50% of states had means above this value; 50% were below
25% of states had means below this value; 75% were above
Only 5% of states had means below this value
This is the lowest state mean in the distribution
O Circles represent states with extremely high or extremely low mean scores, compared to other states; in the top 5% or the bottom 5% of the Distribution
300
290
280
270
260
250
240
230
220
210
All

Graphical displays of distributions of states' mean NAEP TSA scores for important subpopulations of students, such as the box-and-whisker charts shown above, might help to forestall unwarranted generalizations based solely on the overall samples of students assessed in the states.

Proposition 6: *When testing agencies do not provide interpretations of test results, the public and the press will create their own. Occasionally, but not frequently, these interpretations will be accurate.*

How can the NAEP TSA results be explained? Why did students in one group of states exhibit average eighth grade mathematics proficiency that was so much higher than that of students on another group of states? Compared to other states, why were results so good (poor) in a particular state? Although the design of the NAEP TSA does not lend itself to causal explanations, the press typically demands them from those in positions of authority. The newspaper articles of June 7, 1991, illustrate quite well that persons who hold political office or high-level positions in education bureaucracies feel compelled to explain their states' NAEP performances. Explanations offered on June 7th ranged from that of the Governor of North Carolina, who attributed his state's relatively poor showing to "rigid, centrally controlled schools," to Suzanne Triplett, North Carolina's Assistant Superintendent for Research and Evaluation, who, as noted above, explained North Carolina's performance by citing "a lack of parental involvement, too much television, and a poorly focussed curriculum" *(Greensboro Daily News)*, to H. Dean Evans, Indiana's

Superintendent of Public Instruction, who noted that "Indiana rated well compared with other states because the national average was so low," but attributed the relatively poor performance of the state's Hispanic and black students to "the need for minority teachers" *(Indianapolis Star)*, to Wayne G. Sanstead, Superintendent of Public Instruction in North Dakota, who attributed his state's performance to "the rural verities ... The work ethic would be number one--whatever you do, you work at it, including school" *(Washington Post)*.

Although displays of NAEP TSA results cannot supplant the explanations and interpretations that will surely follow publication of any state-by-state performance reports, they can serve to inform such explanations and interpretations. Null results might be as valuable as those suggesting strong relationships between students' average performances and characteristics of the states or their schools, since such results negate potentially causal explanations.

As noted earlier, the NAEP TSA reports released on June 6th were designed to serve the needs of policymakers, and only secondarily to inform the public about the results of the Trial State Assessment in mathematics. The data displays in these reports were therefore in tabular form, a medium that is most useful for providing detailed information concerning individual states' standings on a multitude of variables and detailed results on the components of relational analyses. Tabular displays are of far less value to the public than to policymakers, and as tools for describing underlying trends and global relationships, they are of little value to anyone. Univariate data tables, apart from sports league tables, rarely excite public interest, and cross-tabulations are too complex to serve as effective means for communicating relationships to a lay audience. The Mullis et al. (1991a) Executive Summary contained a condensed verbal summary of relationships between state-by-state differences in mean student performances and a host of contextual variables, but these relationships were neither illustrated nor described in terms of strength or significance (17):

> An examination of contextual background data from the NAEP assessment and other sources suggests that the higher-performing states tended to have had fewer students in large-city schools, fewer students in free-lunch programs, smaller percentages of black and Hispanic students, smaller percentages of students watching six hours or more of television each day, and larger percentages of students with both parents in the home. Higher-

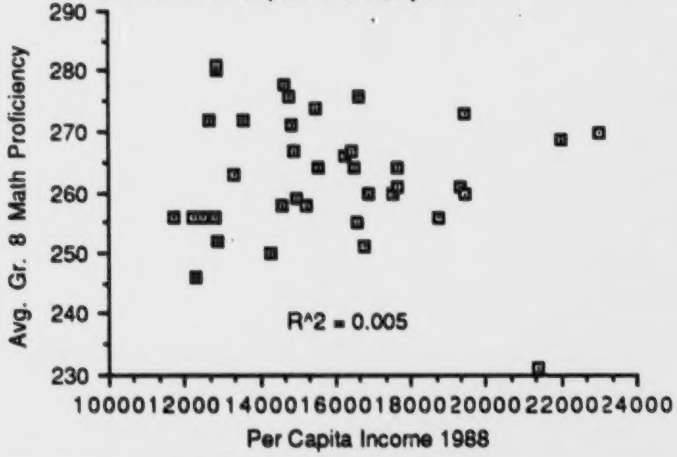
performing states also tended to be less densely populated in general. [This latter finding is curious, in light of the scatter diagram shown below. The lower-performing states tended to be in the Southeast.

It is not surprising that the press paid little attention to these correlates of between-state differences in average performance on the NAEP TSA. The *Washington Post* article of June 7, 1991, offered no explanations for between-state differences; the *New York Times* article of the same date noted only that " ... the highest scores in eighth grade mathematics came from students in the upper midwestern and northwestern states, which in general have few large cities and less poverty than other regions." The article then went on to quote Richard Boyd as saying, "Results show that higher-performing states tend to have fewer students in large cities and fewer students living in poverty." The *USA Today* article of June 7 cited only three correlates of between-state differences in mean student performance, indicating that there were "strong links with scores": "TV. The 10 states where eighth graders watch the least have the highest scores ... Ability grouping. Three of the top states – North Dakota, Montana and Nebraska – put most eighth graders in class without regard to their skills ... Two-parent families. States with highest math scores have the highest percentages of both parents home." Articles in the two local newspapers reviewed (the *Greensboro News and Record* and the *Indianapolis Star)*, which devoted greatest attention to public officials' explanations of their states' results, contained no citations of relational findings from the NAEP TSA reports.

Also noteworthy is the decision of reporters for the three major national newspapers and the two local newspapers reviewed to ignore the press release statements prepared by two members of the National Assessment Governing Board (NAGB) and the Council of Chief State School Officers. One NAGB member prepared a two-and-a-half-page statement devoted principally to observations concerning relationships among between-state differences in eighth graders' average math proficiencies and various demographic characteristics of the states, student background variables, students' perceptions of their mathematics abilities, and sources of school support (public vs. private). The observations of this NAGB member were data-based and informative, although they provided no information concerning strengths of relationships. Perhaps, in the judgment of the reporters, his name and position did not carry sufficient "weight" to warrant citation. The statement prepared by the second NAGB member provided little data-based information concerning the correlates of between-state differences in student performance on the NAEP TSA in grade 8 mathematics. Instead, it echoed the statement of Secretary Alexander concerning eighth graders' abysmal overall performance, provided a set of undocumented recommendations for solving this problem, and concluded with a strongly worded endorsement of the development of achievement levels for NAEP. The Council of Chief State School Officers' press release extolled the virtues and values of NAEP, cited the activities of the Council in support of NAEP and the improvement of mathematics achievement in the Nation's schools, and called for Congressional authorization and Presidential support for extension of the NAEP TSAs to 1994 and beyond. Clearly, these latter statements provided no useful, data-based explanations of state-by-state NAEP results and were summarily ignored by the press.

Graphical displays might better inform the public about variables that are and are not related to state-by-state differences in NAEP TSA outcomes than do the tabular displays currently used by NCES and ETS. In addition, appropriate graphs might forestall explanations by office holders and policymakers that are unlikely to be warranted. Some examples, developed from results of the 1990 NAEP TSA in mathematics, follow.

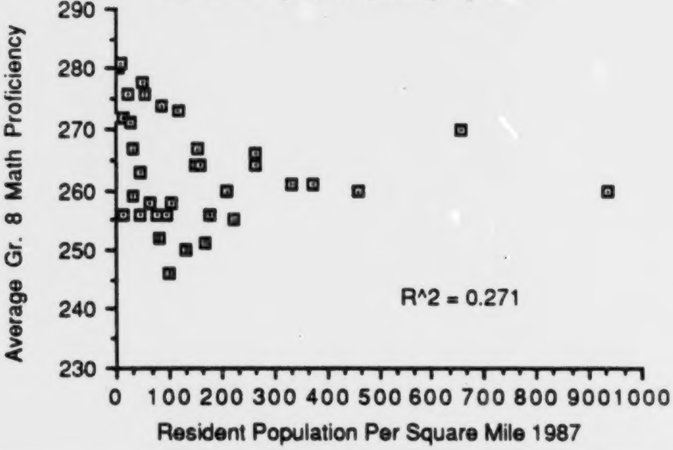
Fig. 4 Relationship Between Average Grade 8 Math Proficiency and 1988 Per Capita Income, by State



This graph provides an example of a null result. It shows that students' average eighth grade NAEP mathematics proficiency is essentially unrelated to the per-capita income of residents of their state. Armed with this data display, members of the public and the press would not accept their state's per-capita income as an explanation of their students' NAEP eighth grade mathematics performances. A similar example is provided by a graph showing the relationship between states' population density and the average

NAEP mathematics performances of their eighth graders. One outlier, Washington, D.C., is not shown on this graph. With that point excluded, there is no evidence that eighth graders living in more (or less) dense states performed better (or worse) on the NAEP TSA in grade 8 mathematics.

Fig. 5 Relationship Between Grade 8 Math Proficiency and 1987 Population Density, by State



In contrast to these null findings, graphs illustrating relationships between students' average performances on the NAEP TSA in mathematics, the average percentage of a state's eighth graders who are eligible to receive free lunch in school, and the percentage of a state's students who are members of a minority group, provide information that might well be explanatory. Our public schools are far less effective in educating minority students and children of the poor than they are in educating children of middle-class majority families.

Fig. 6 Relationship Between Average Grade 8 Math Proficiency and Percent Minority Students in 1986, by State

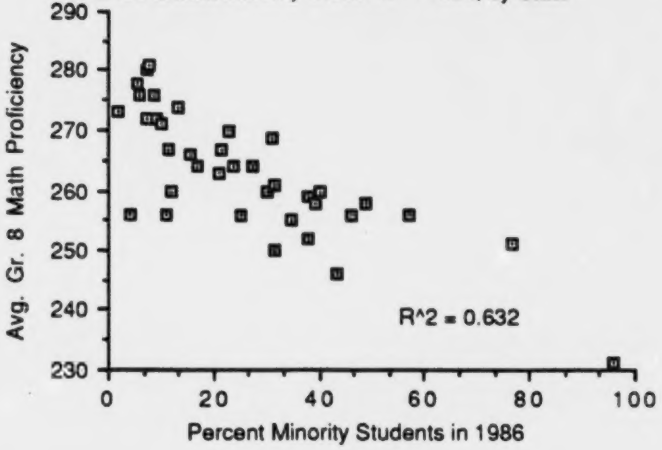
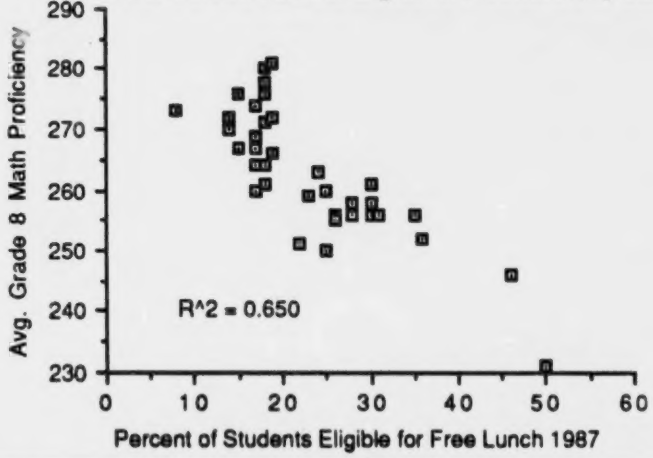


Fig. 7 Relationship Between Grade 8 Average Math Proficiency and 1987 Percent Students Eligible for Free Lunch, by State



These displays illustrate clear and substantial between-state relationships. Over 60 percent of the variation in states' average eighth grade performance on the NAEP TSA is associated with the states' percent of students who are minority group members, and with states' percent of students eligible to receive free lunch in school. These graphs have the advantage of being simple to understand (and are thus likely to satisfy the requirements of Proposition 1), and, if accompanied by appropriate warnings and caveats concerning causal interpretations of correlational data, should be consistent with the requirements of Standards 3.16, 6.8, and 15.10.

Proposition 7: *Reports of NAEP TSA results should not depend solely on the normative properties of the NAEP scale.*

Standards 4.1 and 4.2 have not been satisfied in the 1991 reports of NAEP TSA results intended for public consumption. Standard 4.1 requires that scales used for reporting results and the rationale for choosing them be described clearly, so as to facilitate accurate interpretation of scores. Among other provisions, Standard 4.2 requires that "When scaled scores are used, the scale chosen should be consistent with the intended purposes of the test and should be described in detail."

In a number of the press reports that appeared in the June 7, 1991 newspapers, the descriptions of anchor levels provided on page 6 of the Mullis, et al. Executive Summary were cited:

> Level 200 – Simple additive reasoning and problem solving with whole numbers.
> Level 250 – Simple multiplicative reasoning and two-step problem solving.
> Level 300 – Reasoning and problem solving involving fractions, decimals, percents, elementary geometry, and simple algebra.
> Level 350 – Reasoning and problem solving involving geometry, algebra, and beginning statistics and probability.

The sterility of these "public" definitions of anchor levels and the degree to which they fail to communicate the diversity of problems associated with each of the anchor levels becomes starkly apparent when they are compared with the full definitions shown in Figure 1.1 (56-57) of the full composite report (Mullis et al. 1991b). Such statements as "Only 46 percent of high-school seniors can solve problems involving fractions, decimals and percentages. Only 5 percent are ready for college math," appearing in the July 7, 1991, issue of *USA Today,* misrepresent the NAEP mathematics performances of the nation's students and result, in part, from the sterile descriptions of anchor levels provided to the press. Consider the full description of Level 300:

> Students at this level are able to represent, interpret, and perform simple operations with fractions and decimal numbers. They are able to locate

> fractions and decimals on number lines, simplify fractions, and recognize the equivalence between common fractions and decimals, including pictorial representations. They can interpret the meaning of percents less than and greater than 100 and apply the concepts of percentages to solve simple problems. These students demonstrate some evidence of using mathematical notation to interpret expressions, including those with exponents and negative numbers.
>
> In measurement, these students can find the perimeters and areas of rectangles, recognize relationships among common units of measure, and use proportional relationships to solve routine problems involving similar triangles and scale drawings. In geometry they have some mastery of definitions and properties of geometric figures and solids.
>
> In data analysis, these students can calculate averages; select and interpret data from tabular displays, pictographs, and line graphs; compute relative frequency distributions; and have a beginning understanding of sample bias. In algebra, they can graph points in the Cartesian plane and perform simple algebraic manipulations such as simplifying an expression by collecting like terms, identifying the solution to open linear sentences and inequalities by substitution, and checking and graphing an interval representing a compound inequality when it is described in words. They can determine and apply a rule for simple functional relations and extend a numerical pattern.

That the true expectations of student performance at Level 300 are not adequately conveyed through "Reasoning and problem solving involving fractions, decimals, percents, elementary geometry, and simple algebra" is obvious. Equally obvious is the invalidity of press reports that result from the reduction of full anchor-level definitions to one-line summaries that fail to encapsulate or convey the richness and complexity of student performances associated with these anchor levels. Only the solution to this problem is elusive. The press cannot use the full definitions and few members of the public would understand or appreciate them even if they appeared on the front page of the nation's daily newspapers. Perhaps descriptions of anchor levels prepared for public dissemination could incorporate associated test items more directly: "Most students who perform at Level 300 can solve problems involving fractions and decimals such as this one ... , problems involving measurement such as this one ... , algebra problems such as this one ... , geometry problems such as this one ... , and can read and interpret graphs such as this one ... . The definition would incorporate five example items at a minimum, making it longer than many reporters would desire. But the shorter definitions currently provided fail to convey the diversity of knowledge represented by the NAEP mathematics anchor levels.

As noted earlier, most of the newspaper reports of June 7, 1991, referred to a maximum score on the NAEP mathematics scale of either 500 or 350, implying in either case, that eighth grade students should have achieved these maximum scores. Reporters and the public are used to examinations that are homogeneous in content and appropriate to the ability levels of the groups for which they were prepared. Since it is based on a scale that, by design, spans the expected abilities of students over eight years of public schooling, the NAEP assessment in mathematics is inconsistent with this public expectation. NAEP reports prepared for public dissemination must convey, far more clearly than is currently the case, the unitary, cross-grade nature of the NAEP scale. Reporters must be told that virtually all eighth graders should perform above Level 250, many should perform at Level 300, but that hardly any would be expected to perform at Level 350. The grade-level designations associated with the anchor levels in the Elliott and Phillips (1991) press release (3) did not dispel the notion that appropriately performing eighth graders should have earned the "maximum" score of 500 (or 350). It is not clear from this press release or the Mullis et al. (1991a) Executive Summary (6-7) whether the grade-level designations associated with the anchor levels represented current instructional practice in the United States or the desires of mathematics educators who prepared the National Council of Teachers of Mathematics *Standards*. If the former, the sources of data concerning current instructional practices should be cited. This point is not to be dismissed lightly, since it relates to the issue of whether students in the nation's schools are being taught content that is tested in the NAEP mathematics assessment and not learning that content, or whether the content of instruction fails to include subject matter that students in grades 4, 8, and 12 would reasonably be expected to master. In addition, the inclusion of grade-level designations for the NAEP anchor levels in the press briefing materials is inconsistent with the requirement of *Standard* 4.1 that the derivation of all scaled scores be described in an associated publication (American Psychological Association 1985). None of the technical appendices to the ETS report on the 1990 NAEP results indicates that grade-equivalents were associated with the NAEP mathematics scale in any systematic way.

None of the press materials distributed at the briefings of June 5 and 6, 1991 provided readers with guidance concerning score differences among states that should be considered educationally significant, in contrast to statistically significant. The absence of such information might reflect a fundamental deficiency in the way NAEP TSA results (and, for that matter, national NAEP results as well) are characterized in terms of the NAEP scale.

For example, in statistical terms, Idaho's mean score of 272 on the mathematics TSA is significantly higher than Indiana's mean of 267; the five-point difference in sample means exceeds 3.8 standard errors of the difference. But what should be said of a five-point difference on the NAEP scale, particularly when both values are midway between the anchor points 250 and 300? Is five points indicative of any real difference in average mathematics performance? Or would 10 points indicate an educationally important difference? For that matter, what can be said of students who have a mean score of 272? How does one ascribe meaning to scale scores that fall between scale anchor values? An obvious rejoinder from ETS and NCES personnel might be that readers should ignore state means and concentrate instead on percentages of students in each state whose performance exceeded each of the scale anchor values. Their supporting argument might be that the performances of students at the scale anchors are well described. As noted earlier, the claim that anchor values were well described is dubious, at least in the press briefing materials and the information that appeared in the nation's press following the June 5 and 6, 1991 briefings. But more to the point, the press briefing materials placed great emphasis on average scale values, both as univariate descriptive statistics and in analyses of the correlates of student performance on the NAEP TSA in mathematics. This emphasis on NAEP scale means undoubtedly contributed to the frequently inferred, but erroneous, conclusion that U. S. eighth graders were taking a test with a maximum possible score of 500 (and the implicit conclusion that 500 is what they *would* have earned if the schools were doing their jobs), but they only achieved a mean of 265, suggesting that eighth graders only answered 53 percent of the items correctly.

These observations present a dilemma without an obvious solution. Linn (1990) noted that Ralph Tyler's early plans for NAEP included reports to the public on what Americans at various age levels, living or going to school in particular parts of the country, who were members of various racial or ethnic groups, with families in various economic circumstances, knew. Tyler's strategy for communicating this information was through the presentation of tasks that examinees attempted, together with simple statistics on the percentage of the examinees in each category who completed each task successfully. The . materials and the Executive Summary (Mullis et al. 1991a) distributed at the press briefings of June 5 and 6, 1991 provided some information in this form, but it was very limited. The Executive Summary contained only two NAEP items, one of which was administered to students at all three tested grade levels, the other administered only in grade 12. A set of "Example Items" accompanied the Elliott-Phillips (1991) press release. It contained two items associated with Level 200, three items associate with Level 250, two items associated with Level 300, and only one item associated with Level 350. Several press reports that appeared on June 7th incorporated these items, but they were obviously not sufficient to inform the public on the operational meaning of the NAEP anchor points or to assist the public in interpreting mean scale values that fell between the anchor levels. Additional released items were available to the press in the comprehensive NAEP report. Perhaps the entire set should have been included in the supplement to the Elliott-Phillips press release.

Merely releasing additional NAEP items in a document that calls attention to their availability will not solve the problems of scale interpretation raised above. Even if the public knew quite well what was meant by performances at the four anchor levels, they would be hard pressed to know whether a difference of 10 points between state means on the NAEP scale was worthy of their concern. They also would find difficult, interpretation of means that fell between adjacent anchor levels, such as the two cited earlier. Perhaps a more radical departure from current reporting strategies is needed. Since mathematics is not a unitary subject, and students' development is not constant across grades, NCES and NAGB might attend to the often-voiced suggestion that results be characterized in terms of well described within-grade content-area scales.

To the credit of ETS and NCES personnel, Linn's (1990) suggestion that released NAEP items be accompanied by overall item difficulty statistics (the proportion of students in the grade(s) at which the item was used who answered the item correctly), was heeded. And on June 7th, the only national newspaper that included a NAEP item in its report of results (*The Washington Post*), reported the associated item difficulty statistic correctly. Earlier reports that omitted item difficulty values, but provided *conditional* statistics on percentages of students with NAEP scale scores at the four anchor levels who had answered each item correctly, led to frequent misinterpretation of NAEP results in the public press. One misinterpretation was based on the assignment of released NAEP items to anchor levels, and associated statistics on the percentage of students with scores exceeding each anchor level. For example, the item used by *The Washington Post* in its June 7th report on NAEP results "The weight of an object on the Moon is 1/6 the weight of that object on the Earth. An object that weighs 30 pounds on Earth would weight how many pounds on the Moon?" was associated with Level 300 on the NAEP scale. According to Mullis et al. (1991b, 55), only 14 percent of eighth graders in the national NAEP sample scored at or above Level 300. In the past, some reporters used statistics such as this to draw conclusions equivalent to the statement that only 14 percent of U. S. eighth graders could answer the item correctly when, in fact, 49 percent did so (see Linn 1990, for a more complete discussion of this issue).

A Reprise, a Recommendation, and an Observation

The press reports cited in this paper, in juxtaposition to press release documents and summary reports, provide some insights about reporting strategies that did not convey desired messages to the press, and ultimately to the public. In response to these data sources, in keeping with the 1985 *Standards for Educational and Psychological Testing*, and consistent with some personal conjectures, some alternative formats for reporting NAEP TSA results have been recommended. These recommendations constitute *hypotheses* concerning what the press will accept and use, what the public will understand, and what both will interpret in defensible ways. Finding strategies for reporting NAEP TSA results (or any achievement test results) in ways that accurately convey their psychometric and statistical limitations; provide useful, understandable information to the public; promote defensible interpretations while controlling indefensible ones; and do not violate the constraints imposed by the public press, is challenging to the point of near impossibility.

The press and the public would be well served if these hypotheses, and others suggested by persons responsible for reporting NAEP TSA results, were systematically tested through carefully planned and thoughtfully executed studies prior to their adoption. Several possibilities come to mind. A series of focus group sessions involving members of the press and other media, a cross-section of newspaper readers, and a cross-section of those who gain knowledge of current events solely through television could be held to determine what forms of tabular and graphical presentation of NAEP TSA results participants understood, what reporting formats they found interesting and uninteresting, how they interpreted results presented in various formats, and what formats they preferred.

In a separate study, simulated newspaper articles describing NAEP TSA results could be prepared, and randomly sampled newspaper subscribers could be asked to describe what they learned from the articles and how much interest the articles held. Such a study would provide information on preferred reporting formats, the relationship (if any) between reporting formats and the likelihood of defensible interpretations of NAEP TSA results, and reporting formats that were totally ineffective in conveying understandable information about NAEP TSA results. A carefully planned experiment could be designed to balance tabular versus graphical portrayal of test results, short versus long verbal interpretations of results, simpler versus more accurate descriptions of statistical and psychometric caveats, etc. A parallel experiment could be used to learn about the construction of effective television reports of NAEP TSA results.

The NAEP project staff at Educational Testing Service and in the National Center for Education Statistics worked hard to develop press materials that conveyed the essence of NAEP TSA results fully and accurately. They held extensive press briefings and made available to the press, full reports, knowledgeable experts on the National Assessment, and report summaries. Apparently, the constraints of publication deadlines and reporters' needs to attend to competing demands interfered with effective communication of NAEP TSA results and limitations to members of the press. To ameliorate this problem in the future, those responsible for the NAEP TSA (and "mother" NAEP as well) should host a series of working meetings with the press beginning several weeks prior to release of the NAEP results. These meetings should be structured around the implications of NAEP TSA findings, alternative ways of conveying those implications, important caveats and limitations imposed by NAEP TSA methodology, and defensible (and indefensible) interpretations of NAEP TSA results. Possible suggestions that the government was trying to dictate the reporting of NAEP TSA results notwithstanding, portions of simulated newspaper articles containing data displays and defensible interpretations might be distributed for discussion. If these activities took place during a series of workshops rather than just one, members of the press could be interviewed to determine what information they found most useful for their articles, how they interpreted that information, and what they preferred.

If successful, the results of these studies could be used by the NAEP staff at Educational Testing Service and NCES to prepare a short guide for state policy-makers on reporting NAEP TSA results to members of the press and their constituencies. The potential for improving the accuracy and utility of NAEP TSA reporting is obvious.

A final observation on public reporting of NAEP results concerns informal, rather than formal, reports to the public. One cannot help but be struck by the consistently negative tone of the the statements concerning the 1990 NAEP results issued by the Secretary of Education and the Chair of the National Assessment Governing Board. The message delivered was one of gross failure in mathematics on the part of U. S. students, and by implication, failure on the part of the nation's public schools. Whether this was an accurate and defensible message will be debated by the nation's educators for months to come. But regardless of its validity, the message was immediately echoed by the press in its reports on 1990 NAEP mathematics results.

Although no state's eighth graders exhibited NAEP mathematics performance that was laudable, and relatively small percentages of eighth graders could handle test items associated with NAEP Level 300, virtually all eighth graders in the participating states performed well on the items associated with Level 200, and a majority of students in almost all states performed well on the items associated with Level 250. There is no doubt that this finding warrants serious attention on the part of school boards, school administrators, and mathematics educators throughout the nation. However, the NAEP findings do not support the conclusion that the public schools have failed or that the public schools are incapable of teaching a large percentage of eighth graders to solve Level 300 mathematics items. Considering the press for basic skills instruction and the accountability-driven education policy that has constrained the curricular choices of U. S. schools in recent years, it is not clear that they have *attempted* to teach eighth graders the kinds of skills needed to solve Level 300 test items. This basic skills mathematics curricula of U. S. public schools is responsive to legislated minimum competency testing programs in the vast majority of states. The Secretary might have speculated that the schools were doing what they were told by state boards of education and state legislatures, but that what the schools were told was shortsighted and ineffective.

The statements by the Secretary and the NAGB Chair raise fears that NAEP has been transformed from the barometer of subject matter competence and achievement envisioned by its founders, to a mechanism for attempting the reform of public education and the promotion of an as-yet-unclarified political agenda. This transformation might well affect the long-term viability of NAEP as a valid longitudinal indicator of student achievement in the nation. To the degree that reporting of NAEP TSA results (whether through formal reports or the informal comments of high-level federal policymakers) raises the stakes associated with states' NAEP performances, the results will become less trustworthy. Evidence based on state legislatures' attempts to impose high-stakes achievement tests indicates that such tests result in narrowed curricula, superficial instruction, and (most relevant to the current topic) assessments of student achievement that have little generalizability beyond the specific content and format of items that compose the tests (see the November 1991 issue of the *Phi Delta Kappan* for a series of papers on this point).

Reporting to Policymakers

Reports of 1991 NAEP TSA results for state-level policymakers were designed over a period of months as a result of extensive discussions among members of a 10-person Analysis and Reports Committee and extensive negotiations with members of the NAEP Trial State Assessment Network composed of directors of testing, state mathematics coordinators, and state mathematics supervisors in all participating states (Owen 1991). The resulting comprehensive report and individual reports for states contain a wealth of data that have been organized, displayed, and described in a logical, well organized sequence.

Several of the observations advanced in the section of this paper that addressed reporting to the public apply with equal force to reports prepared for policymakers. In addition, a number of the 1985 *Test Standards* are equally applicable to reports prepared for policymakers. Standard 2.10 (standard errors of measurement should be reported), Standard 3.16 (report forms should facilitate appropriate interpretations), Standard 4.1 (scales and their rationale should be described clearly), and Standard 4.2 (scales should be consistent with intended purposes) fall into this category.

Standard 2.10 has been satisfied in all of the data displays contained in the composite report and the reports prepared for state-level policymakers. Each estimate of students' mean NAEP performance is accompanied by a standard error that reflects components of variance due to sampling of students and sampling of test items. Although these estimates do not reflect variation that would be present in a multiple-administration assessment of measurement reliability, they incorporate variance components that would be present in single-administration estimates. Reports for policyakers also satisfy the requirements of Standard 4.1. Development of the NAEP proficiency scale is summarized on pages 455 and 456 of the Mullis et al. composite report (1991b) and a brief rationale for the technical development of the scale is provided as well. References to more detailed reports are also provided.

Whether the 1991 NAEP reports satisfy the requirements of Standard 4.2 is questionable. Although developers of the scale would argue that it is consistent with the purposes of NAEP, for the reasons enumerated in the previous section of this paper, their contentions could be challenged: Use of a cross-grade scale constrains within-grade variation and facilitates misinterpretation of the scaled performances of students within a grade. The meaning of scale values that fall between described anchor levels is unclear. Differences in states' mean performances can be described readily in terms of their statistical significance, but with great difficulty (if at all) in terms of their educational significance.

Apart from their satisfaction of the requirements imposed in the 1985 *Standards for Educational and Psychological Testing*, the utility of the current NAEP composite and state-level reports is an empirical question. Although ETS and NCES have "given state policy-makers what they asked for," the reports contain an overwhelming amount of largely undigested data. Narrative summaries of data tables were often telegraphic, concentrating on general trends, with brief allusions to states that deviated from the trends. However, strengths of relationships were rarely described and never portrayed. The reports leave to the reader the task of gleaning accurate mental pictures of cross-state relationships from the scores of numbers provided in each of dozens of data tables. As a case in point, consider the narrative summary that accompanies the 16 data tables dealing with teachers' instructional emphases in eighth grade mathematics classes in Chapter 15 of the composite report (355-356):

> Across the states participating in NAEP's 1990 Trial State Assessment Program, teachers reported the greatest instructional emphasis in two mathematics content areas--numbers and operations and algebra and functions. The results suggest that relatively few eighth graders attending

> public schools were provided the opportunity for in-depth learning in measurement or geometry, and very few were given the opportunity for in-depth learning in data analysis, statistics, and probability.
>
> It appeared that the more mathematically proficient eighth graders were receiving considerable instructional emphasis in the area of algebra and functions, whereas the less-proficient eighth graders appeared to be given little or no opportunity to learn this subject and were receiving heavy instructional emphasis in numbers and operations. Nevertheless, there were some variations from state to state. For example, 40 percent of the eighth graders in New York were receiving heavy instructional emphasis in geometry.
>
> For eighth graders attending public schools, regardless of the content area emphasis--numbers and operations as opposed to algebra and functions--teachers reported emphasizing learning facts and concepts as well as procedures more than learning reasoning and communication skills. This finding was pervasive, although teachers in some states reported a somewhat more balanced approach, particularly in the District of Columbia, and to some extent in California, Colorado, Maryland, and Oregon.

This summary is factual. However, it does not communicate the degree of consistency or diversity in states' adherence to the trends it describes, and provides little information on states that "bucked the trends." The kinds of graphic displays illustrated in the section of this paper concerned with public reporting of NAEP results would likely be of value to policymakers as well, since they would facilitate communication of strengths of relationships, numbers of anomalous states, and (if individual data points were identified in displays) the identity of anomalous states.

Although NCES and ETS were sensitive to the perceived needs of state-level student assessment personnel and state-level policy-makers in their organization and preparation of reports on the 1990 NAEP TSA (as well as their reports on "mother NAEP") and applied the fundamental NAEP consensus philosophy to cooperative planning of the content and organization of the state reports, the ultimate utility and use of these reports must be examined carefully. The extensive planning sessions conducted by NCES resulted in "dummy" reports containing "table shells" that corresponded to each table in the final state reports (Table shells have the format, column headings, row headings, titles, etc. of actual data tables, but they do not contain real data). State assessment personnel agreed to the content and organization of the state reports on the basis of these dummy reports and table shells. Now that these educators have had actual NAEP TSA results for their states in the agreed-upon format for some time, NCES and NAGB should ensure that studies of the utility of these reports are conducted.

Consonant with Messick's conception of the consequential elements of content validation (1989), it is also necessary to determine whether the states' applications of NAEP TSA reports have been responsible and consistent with the kinds of inferences these reports can reasonably support. In sum, it is essential that those responsible for future NAEP assessments better understand state policy-makers' uses of their reports and, following a period of use, the strengths and deficiencies of current NAEP TSA reporting strategies, reports, and data displays.

To ascertain and evaluate the behaviors and the experiences of NAEP TSA report users, NCES, NAGB, and their NAEP contractors should support a well planned evaluation that includes (1) interviews of state directors of testing, state mathematics coordinators, and state mathematics supervisors concerning their use of their state's NAEP TSA report and the NAEP comprehensive report; (2) content analyses of state-developed reports on NAEP TSA results and the policy recommendations (if any) derived from NAEP TSA results that have been issued by agencies of state governments since June 6, 1991; (3) interviews with state legislators who are members of committees responsible for K-12 education to determine what they have been told about their state's NAEP TSA results, what actions, if any, are contemplated in response to the results, and the degree to which they found the results to be useful; and (4) content analyses of a sampling of newspapers in each state and territory that participated in the 1990 NAEP TSA, to determine whether derivative public dissemination of information on the NAEP TSA took place and the soundness of such information.

Although NAGB has encouraged the Congress to institutionalize state-by-state NAEP reporting on a continuing basis, the Congress has wisely elected to subject state-by-state NAEP to an extended trial period. Initial evaluative information assembled by the National Academy of Education panel indicates that the operational elements of the NAEP TSA have been largely successful: almost three-fourths of the states elected to participate; within participating states, rates of participation of schools were reasonably high, as were student response rates; data were collected, analyzed, and reported in a timely fashion. Although this outstanding record of achievement should earn plaudits for the NAEP staff and its contractors, procedural success is insufficient to warrant the institutionalization of state-by-state NAEP. Congress should act on NAGB's recommendation only after an evaluation demonstrates the usefulness of state-by-state NAEP to the formation of national education policy, the formation of state education policy, or that the availability of state-level NAEP data facilitates the improvement of education in participating states, and that the availability of state-by-state NAEP data has beneficial consequential effects. An investigation of the cost/benefit relationship of state-by-state NAEP results might also serve to inform sound Congressional action. If well planned and conducted, the evaluation studies recommended earlier should provide this essential information.

Conclusions and Recommendations

Despite the substantial and first-rate work of NCES and ETS professionals responsible for NAEP, reports intended to facilitate public communication of NAEP TSA results (the Mullis et al. *Executive Summary* and various press releases) did not realize their public reporting objectives. Newspaper reports of NAEP TSA results published the day after the NCES/NAGB press briefings were rife with misinterpretations and erroneous conclusions. In addition, it could be argued that NAEP TSA reports prepared for public consumption did not satisfy the requirements of several of the 1985 *Standards for Educational and Psychological Testing*. Reports intended to communicate NAEP TSA results to state-level decisionmakers were factual and well organized. Although it appears that their communicative value could be increased, the usefulness of these reports is an empirical question. This question will be answered, in part, through follow-up studies currently planned and being conducted by the NAE Panel on the NAEP Trial State Assessments and, perhaps, through the additional studies recommended in this paper.

Recommendations for improvement of the reporting of NAEP TSA results are as follows:

1. NCES and ETS should conduct a series of workshops for media personnel prior to the release of future NAEP results for the purpose of educating such personnel about defensible and indefensible

interpretations of NAEP findings, and to determine the types of data reports and displays that best satisfy the reporting needs of the media and their audiences. Additional studies of data reports that facilitate or inhibit the public's understanding of NAEP TSA results should also be conducted.

2. The utility of NAEP TSA reports intended for use by policymakers should be assessed through follow-up studies. In the interim, inclusion of data displays that portray strengths of relationships as well as anomalous cases should be considered for reports of the 1992 NAEP TSA.

3. Public reports of NAEP TSA results should contain interval estimates of states' mean performances rather than point estimates.

4. Overgeneralizations of NAEP TSA results by the press should be countered by including specific warnings and examples of interpretations that *are not* supported by results in public reports of NAEP TSA results. Examples of interpretations that *are* supported by results should be contrasted with examples of those that are not.

5. Public reports of NAEP TSA results should include simple graphs that emphasize the overlap and uniqueness in states' distributions of scores. Data displays and tables should call attention to the variability, as well as the central tendency, of states' score distributions.

6. Relationships between NAEP TSA results and characteristics of states and students should be summarized in simple graphical displays that portray strengths of relationships as well as anomalous cases.

7. Effective strategies for describing the meaning of states' NAEP TSA means that fall between NAEP anchor levels should be developed and disseminated.

8. Effective strategies for identifying and describing the educational significance of differences between states' NAEP TSA means (and

other distributional characteristics) should be developed and disseminated.

9. Public reports of NAEP TSA results should include richer descriptions of the meaning of student performance at the NAEP anchor levels and larger numbers of example test items associated with each anchor level.

References

Alexander, L., and James, H. T. 1987. *The Nation's Report Card: Improving the Assessment of Student Achievement.* (Report of the Study Group). Washington, DC: National Academy of Education.

Ambach, G. May 17, 1991. Personal communication with the author.

American Psychological Association. 1985. *Standards for Educational and Psychological Testing.* Washington, D.C..

Elliott, E. J., and Phillips, G. W. .June 6, 1991. *The State of Mathematics Achievement: NAEP's 1990 Assessment of the Nation and the Trial Assessment of the States.* Press release. Washington, D.C.: National Center for Education Statistics.

Haertel, E. H. 1991. "Should the National Assessment of Educational Progress Be Used to Compare the States?" *Educational Researcher, 20:* 17.

Koffler, S. L. April 1991. "Reporting Results of NAEP's Trial State Assessment Program. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Linn, R. L. October 26, 1990. "Historical Origins and Issues in the National Assessment of Educational Progress." Paper presented at the Institute for Practice and Research in Education Forum entitled "Assessment at the National Level," University of Pittsburgh, Pittsburgh, PA.

Messick, S. 1989. "Validity," in *Educational measurement* (3rd. ed.). Ed. R. L. Linn. New York: Macmillan: 13-103.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G. W. 1991a. *The State of Mathematics Achievement: Executive Summary.* Report No. 21-ST-03. Princeton, NJ: Educational Testing Service.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G. W. 1991b. *The State of Mathematics Achievement: NAEP's 1990 Assessment of the Nation and the Trial Assessment of the States.* Report No. 21-ST-04. Princeton, NJ: Educational Testing Service.

National Academy of Education. 1991. *April 1991 Interim Report on the Evaluation of the 1990 NAEP Trial State Assessment.* Washington, DC: Panel on the Evaluation of the NAEP Trial State Assessment Project.

Owen, E. June 12, 1991. Personal communication with the author.

Phillips, G. W. 1991. "Benefits of State-by-State Comparisons." *Educational Researcher, 20:*17-19.

Schmidt, W. H., Porter, A. C., Schwille, J. R., Floden, R. E. and Freeman, D. J. 1983. "Validity as a Variable: Can the Same Certification Test Be Valid for All Students?" *The Courts, Validity, and Minimum Competency Testing*, Ed. G.F..Madaus. Hingham, MA: Kluwer-Nijhoff.

U. S. Department of Education, National Center for Education Statistics. 1988. *Digest of Education Statistics*. Washington, D. C.

U. S. Department of Education, National Center for Education Statistics. 1990. *The Condition of Education 1990, Volume 2, Postsecondary Education*. Washington, D. C.

U. S. Department of Education, National Center for Education Statistics. 1991a. *The Condition of Education 1991, Volume 1, Elementary and Secondary Education*. Washington, D. C.

U. S. Department of Education, National Center for Education Statistics. 1991b. *The Condition of Education 1991, Volume 2, Postsecondary Education*. Washington, D. C.

U. S. Department of Education, National Center for Education Statistics. 1991c. *College-Level Remedial Education in the Fall of 1989*. NCES Report 91-191. Washington, D. C.

The Case for District- and School-Level Results from NAEP

Ramsey Selden
Council of Chief State School Officers

August 1991

Background

The past year has seen the advent of one of the most dramatic developments in assessment and accountability in American education--the extension of the National Assessment of Educational Progress to produce state-level results. Reflecting a major turnaround in attitudes toward accountability, this change brought a new level of information about the accomplishments of the education system, and it brought a host of concerns as well. Major concerns were that comparing states would not be constructive or useful, that it would result in distortions of educational practice, and that it could "contaminate" the national NAEP, interfering with its ability to fulfill its original purposes as a national indicator.

The state-level NAEP has been implemented over the past five years with a good deal of consideration. A careful approach for its implementation was negotiated and established by Congress, including the evaluation being conducted by the National Academy of Education (NAE). Part of the political solution of how to implement state-level NAEP was to put school or district-level use of NAEP in limbo. While the state-level program was tried and evaluated, use of NAEP at the school or district level was precluded, presumably so adverse effects or other consequences of the expansion of NAEP could be headed off and not expanded uncontrolled among numerous schools and districts around the country.

Meanwhile, the question of whether NAEP *should* be used at the school or district level has not been answered. Several states have lobbied for it. Most organizations representing local school systems or practitioners oppose it, while a university (the University of Tennessee) organized itself to offer a school- and district-level assessment program which was "NAEP-like," or equivalent to NAEP. But, nobody responsible for NAEP has really explored and attempted to resolve the question of whether NAEP should be used at the local level.

In this paper, the arguments for using NAEP at the local level will be explored. First, some major arguments for school- or district-level use of NAEP will be put forth. Second, some of the potential problems in using NAEP at this level will be acknowledged and considered. Finally, to the extent possible, these potential problems will be weighed in relation to the prospective benefits.

Before we proceed with this discussion, some concepts and parameters for the paper must be discussed. First, this paper will *not* explore using NAEP to produce scores for individual students. That is an important limitation in scope. A student-level NAEP is a profoundly important concept that should be explored very carefully, but that will not be done here. We are talking only about use of district averages from NAEP to examine school or district performance. Second, to the extent possible, the paper will explore both comprehensive implementation of NAEP among schools or districts--as in a statewide assessment program--as well as uses by isolated schools or districts on their own.

Arguments for School- and District-Level NAEP

Let us start by putting forth some of the major reasons for compiling data through NAEP at the school or district level.

Reason 1: Having information is better than not having it.

This argument is so basic as to seem almost facetious, but it must be taken seriously. A reason frequently put forward against collecting information is that having it will result in more harm than good. This reason has been put forth often as a reason for not collecting many types of educational data. The specific harm-benefit issues surrounding local use of NAEP will be addressed later, but here I would like to make a general argument for the value of information. Most of the time, information is useful, and the more of it we have, the better, as long as the information is organized and presented in a way that renders it useful. There are instances when it can be argued that it is better that someone not know something, but we should be clear and careful about these instances. The burden rests on me to argue later that a local-level NAEP would not be so flawed and misleading as an indicator as to be dangerous. But meanwhile, I contend that the prospect of information that is not now available is desirable.

Why do people oppose the collection of more or better information? Mainly, they are opposing change in the information landscape, and there are two reasons for this--one sincere and one cynical. The sincere reason is fear that information will be misinterpreted, misused, or that unfortunate, unforeseen behavior will result from it. The cynical reason is that the new information will make someone look bad or require effort someone does not want to take. We will come back to these later.

When the planning was being done for the 1992 NAEP reading assessment, which was to be done state-by-state on a trial basis, one of the reasons put forth by the International Reading Association in opposing the assessment was that state-by-state comparisons should not be made. (Since then, IRA has adopted a very thoughtful and constructive resolution on large-scale assessment.) (IRA 1991) The implication of the earlier position was that the comparisons among states would be inherently problematical--that it would be educationally wrong or dangerous to produce these comparisons. Similar reasons were put forth early on by state assessment directors to question state comparisons. Meanwhile, the member-delegates of IRA who voted for this position (and the state assessment directors) were custodians of local and state testing programs which compared districts, schools, and students with one another and with ephemeral national percentiles and grade equivalents. It seems more legitimate to compare the empirically estimated performance of entities like states or districts, which have an identifiable responsibility for, and program for conducting, reading instruction, than to judge performance in terms of national percentile ranks. At least one can go see what the state or district across the river is doing and judge apparent performance differences in relation to apparent program differences.

Reason 2: Important, useful information can be generated through NAEP at the school and district level.

As a broad-based data collection program, NAEP offers information with two great assets--its content and its comparability. In content, NAEP generates student performance data, student and teacher background data, instructional information, and school data. As a broad-based program conducted equivalently across many sites, NAEP offers the power of producing information that can be interpreted in relation to comparable information for other sites.

Student achievement data from NAEP is remaining close to the U.S. state of the art in most subjects. Subject-matter frameworks represent professional and political consensus, and they incorporate the state-of-the-art in theoretical thinking about performance in the subjects.

Concerning **political consensus**, as NAEP has become a more high-stakes accountability assessment, its content has had to reflect a more thorough consensus of states, local school systems, and various professional stakeholder groups. Comparisons could not be made if the content were patently different from what many states were doing, from what many local curricula reflected, or from prominent professional views.

As a **professional consensus**, NAEP has, in several subjects, reflected the resolution of major professional differences in its approaches to measuring performance outcomes. This is true in mathematics (where the assessment has been planned with consideration of NCTM [National Council of Teachers of Mathematics] *Standards*), reading (where the framework *was* the crucible for resolving important professional differences), and writing (where the assessment has changed to reflect substantial developments in the ways educators view the writing process). (CCSSO 1988, 1990; NAGB 1990)

How students learn and perform in the subjects assessed in NAEP is constantly being researched. Reading, science, mathematics, writing, English, and history have been the targets of research into their learning and performance by psychologists, linguists, anthropologists, and others in other disciplines. This research has resulted in powerful insights into performance in these areas: the roles comprehension and metacognition play in reading, stage processes in writing, problem solving in mathematics. These insights have illuminated how students understand and perform in the subjects, so they can be assessed better, and NAEP's frameworks and objectives have reflected the progress made through this research.

The **methods** used in NAEP assessments have stayed near the state of the art in most subjects, at least in recent years. An appropriate premium has been placed recently by the field of education on the *validity* in assessment, demanding that new techniques be developed to measure content knowledge and performance more validly. This is a major impetus behind the move toward performance, portfolio, and other new forms of assessment. In 1990, NAEP pioneered the use of calculators in large-scale mathematics assessment and incorporated open-ended items to measure reasoning and problem-solving abilities, following recommendations for the assessment developed in 1988. Prior to this, NAEP's mathematics assessment was wholly multiple choice.

In 1992 the writing assessment will allow more time for students to write and will use a modified prompt procedure, so students can apply a multi-stage process approach when they write, which they have been taught to use in many states. Prior to this change, students were given less time, and the assessment procedures encouraged a one-step writing process.

Also in 1992, the reading assessment will reflect the national state of the art in reading assessment--using long, authentic reading activities, tapping comprehension through open-ended exercises, and exploring portfolio techniques to base measurement on what students produce themselves, while compiling information on classroom instruction activities.

With these kinds of development, NAEP offers potential local users methods of assessing student performance in the core subjects that reflect national consensus content (more later on why this is appropriate); contemporary, research-based views of the subjects; and increasingly innovative methods to validly measure target knowledge and capabilities. The potential achievement information is as attractive as most, if not all, "packaged" large-scale assessment options available to schools or districts, surpassable only by a well-done, custom, local program.

In addition to student performance data, NAEP offers other kinds of data. Student background variables allow analysis of performance patterns for various demographic groups and in relation to student and household characteristics. For example, NAEP reading data can be analyzed by race-ethnicity, sex, type of community, and in relation to reading practices of students and their families. This has shown, in the past, at least associations between reading performance and the extent to which students are read-to.

Teacher questionnaires yield data on teacher characteristics and instructional practices. The 1990 mathematics assessment revealed important, intriguing relationships between student performance in mathematics, teachers' study of mathematics content in college, and teachers' involvement in professional development activities in mathematics. This is one of the first times large-scale studies have found strong associations between student performance and such teacher variables.

Other variables are collected at the school level, including the presence of special instructional programs and the nature of school management approaches.

Few local assessment programs include these additional kinds of information beyond student achievement. Most local programs stop with achievement and student demographic data. These additional program data are potentially very important to inform planning and analysis of instructional strengths, weaknesses, and needs. NAEP's background data could provide very important diagnostic information, allowing a school or district to describe its teacher workforce, its instructional practices, or the distribution of instructional resources to various student groups, either on their own or in relation to student achievement data. This could be tremendously valuable information to local program analysts and managers.

Not only is the information potentially useful, for local schools or districts to develop this information on their own would be difficult, expensive, and wasteful. If similar kinds of data are needed in most sites, a program like NAEP that offers well-conceived program data to localities could provide a great service.

(It must be noted that whether NAEP now offers "well-conceived" data collection practices is debated. In the past, developing and analyzing effective background questions was difficult for the project. Too many variables were collected, too few analyzed, and many failed to get at important program issues. Great strides have been made recently in focusing on important issues that can be tapped accurately, but further development in this area would still be useful. Meanwhile, it is asserted here that many of those questions currently used in NAEP are important and effective enough to be useful to localities.)

A major feature contributing to the potential value of NAEP to localities is its ability to provide *comparable* data. Descriptions based on one's own data are useful in themselves, of course, but limited. Information acquires more utility by whole dimensions when it is collected over time and when it can be compared across sites. Used by schools or districts, NAEP could be collected regularly, offering trend data on both student performance and background variables. Thus, the school or school system could monitor its progress over time, not only in achievement, but also on background variables: are instructional practices changing in the desired directions? Are the characteristics of the teacher workforce getting better? Are home reading practices improving, perhaps in response to programs to encourage parents to read more to and with their children?

NAEP also offers the tremendous benefit of comparable data from numerous sites. By using NAEP or participating in a statewide program based on NAEP, not only could a school or district know how it is doing, but it could compare its performance with other specific schools or districts, with averages for its type of district, or with results for subgroups of students.

The value of this comparability is obvious for student performance data, but it is equally important for background data. In diagnosing its conditions, a school or district could look at its data on instruction in relation to national averages, state averages, averages for similar types of school or district, or results for specific schools or districts. Is the extent to which we are emphasizing comprehension in reading comparable to other districts around the state or country? Do our students watch television about as much, less than, or more than students in other areas? Are our teachers as well qualified?

The benefits of such comparable data would truly be realized if local collection were done with strong involvement by states, setting up statewide programs to administer NAEP and report it out by schools and districts. That way, each school or district could analyze its data in relation to the full set of schools and districts in the state. To the extent states participated in this way, they could compare themselves with schools and districts around the country and with groups of schools and districts with many organizing features.

Educational planning and management in the United States is done in an information vacuum--virtually in the absence of program data from large-scale data collection. Test scores, fiscal data, and staff evaluations are used, but data on programs is scarce: what is taught? By whom? With what level of effectiveness? To which students? With what resources? And, how do we compare on these matters with other school units? The prospect of this kind of information to support analysis, planning, and management at the local level is tremendous.

Reason 3: Educational data in the U.S. should be vertically integrated where possible--that is, similar at the local, state, and national levels.

Because of its history as a decentralized system, the data collection programs in education in the U.S. have evolved as a combination of local, state, and national programs. Some would argue that, as a result, data collection programs have emerged as a redundant, inefficient hodgepodge of programs. As late as 1986, states counted schools according to 11 different definitions. They counted student enrollment by more than a dozen different procedures. They counted dropouts, if at all, with virtually unique methods. (CCSSO 1986)

In the area of student achievement testing, a 1985 study by UCLA (Burstein et al. 1985) found that states used essentially unique combinations of instruments, grade levels tested, and timing in their testing programs, making it impossible to compare any two states even in reading and mathematics. Due to conflicting goals for state and federal programs, states may have high-quality statewide assessment programs, while local school districts must continue to use commercial, standardized programs as well, for Chapter 1 evaluations.

In the U.S. right now, *different* testing programs are operated at the following levels:

- teacher-made quizzes, tests, and projects;
- tests included in or with published text materials;
- diagnostic and special purpose tests used at the school level, such as those used by school-based reading specialists;
- commercial or custom-developed tests used by the district;
- tests used for program evaluation, which are other than the district-wide or statewide achievement tests;
- statewide assessment programs, which may be commercial, standardized tests or custom-developed programs;
- the National Assessment of Educational Progress, for national monitoring purposes and for state-by-state comparisons; and
- two different international programs for comparing nations and studying achievement phenomena on an international basis--the International Association for the Evaluation of Educational Achievement and the International Assessment of Educational Progress.

In Connecticut, it was found a few years ago in one school district that a dozen different standardized tests in reading were used for various purposes and that it was technically possible for a student to be tested on all of them. (Pecheone 1986)

How can this balkanization and redundancy be reduced? By using the same or equivalent testing programs or data collection programs for more than one purpose. We must structure relationships among testing programs so that when a student is tested once, the results feed a number of purposes up and down the system. We can leave aside the question of whether large-scale testing programs should be used to meet classroom needs; perhaps they can, but we can make stupendous gains in efficiency without tackling that issue. We can have one testing program and one integrated program for collecting other kinds of data serve needs at the local, state, national, and international levels. It would work in the following way: a large-scale assessment battery given to students, perhaps to a sample, would be aggregated to produce school-level results. These would be aggregated to produce district-level results. Through cooperative arrangement with the NAEP program, the same data would be aggregated to generate state-level results (eliminating sampling error in the state data). The state results would be aggregated to produce or contribute to national results. Again, if virtually all states were involved, sampling error at the national level could be eliminated. Finally, by linking or equating NAEP to international programs, school, district, or state results could be converted to international metrics and compared to international results.

What are the advantages of such a system?

- A single test administration could meet the needs of the school, district, state, and nation, eliminating two or three layers of testing, at least.
- Results reported to educators and the public would be expressed in the same metrics and formats; we would be talking about testing in a single, common set of terms.
- Related data would also be comparable and expressed in common terms up and down the system, so our analyses of instructional practices and resources would not be hampered by noncomparable data collection formats.
- If NAEP or NAEP-equivalent assessment became the common metric and came to be regarded as general purpose achievement tests for program assessment, many standardized tests used only for program evaluation could be eliminated.

What programs should be used for such a purpose? They should be valid, high-quality, comprehensive achievement tests measuring subject-matter content and performance that is a true consensus of what students should know and be able to do. Thus, if such vertical integration is to be achieved in student testing, a program like NAEP should be the basis.

At least two models or approaches for implementing such a system in the student achievement testing area are possible: extending NAEP as a single program to many sites or creating a system where variable programs are used that are equivalent to one another for reporting generally and comparatively on student achievement. Here NAEP would be the calibration standard or criterion. There are big differences in the downsides to these approaches that will be discussed later.

For collecting background data, nothing would make sense other than standardizing around a universal set of questions, but volunteerism and local options would be desirable.

Educational data in the U.S. is seriously afflicted by the inability to interpret results from one site in relation to those of another. The creation of state NAEP is one major step in alleviating this problem. Extension to the local level would be another.

Reason 4: We are arriving at national consensus on the goals and content of education.

A serious reason for not using NAEP at the local level was that it would lead to a national curriculum. This reason is valid only if a national curriculum does not exist and is not desired.

As this is written, a National Council on Education Standards and Testing is being convened to study the feasibility of national standards for education. This move reflects a political consensus that some national focus of attention on the content of, and standards for, instruction is desired.

Since *A Nation at Risk* was published in 1983, considerable evidence has accumulated that the achievement of American students is far from desirable. Particularly, few students learn to read, write, or think well, or to handle the deeper, more complex aspects of subjects such as mathematics or science. They do not even seem to be mastering basic facts on literature, history, or geography. They seem adept only at routine, low-level operations in reading, writing, and mathematics.

The evidence also is that decentralization of authority for curriculum and instruction in the U.S. has generally resulted in instructional content which explains why students are learning so little. In mathematics, reading, and science, available evidence suggests an over-emphasis on low-level and decontextualized skills. New materials, assessments, and teacher training programs are correcting this slowly, but 150 years of a decentralized public school system has not resulted in a pervasive, diverse set of strong local instructional programs bringing students, through different paths, to impressive levels of achievement, even to impressive levels of *different* kinds of achievement.

Aside from the argument as to whether we *ought* to have a somewhat centralized curriculum, there is the fact that relatively little diversity *exists*. Due to market forces in educational materials, testing, and teacher training, the content of instruction is not only low level, fragmentary, and content thin, it is consistently so. It cannot be argued that the centralization of effort in education would confront and conflict with a great deal of local diversity. Perhaps more diversity (and diverse forms of high-quality instruction) should exist, but they do not.

So decentralization has not resulted in a great deal of diverse, high-quality, local instruction. American education is already quite homogeneous, but mediocre, and there is political consensus that some national focus is desired. In this environment of interest in national goals and standards, extension of a local NAEP, which represents some commonality of interests in instructional emphasis, seems an appropriate way to proceed.

Liabilities in a School- and District-Level NAEP

What are the downsides to extending NAEP to the district and local level?

1. Use of NAEP in this way could discourage state and local innovation and development in assessment.

It can be argued that, if NAEP were made available at the local level, statewide testing programs and individualized local programs would become as rare as custom-bodied automobiles. This is plausible. Assessment programs are costly, labor-intensive, and time-consuming to develop. Vermont, Maryland, California, Kentucky, Connecticut, Arizona, Michigan, and Illinois can attest particularly to that. Reflecting the era of distinct curricular emphases of the states and districts, these programs were motivated by the need to measure those distinct curricula directly and validly. The assessment development efforts also had the purpose of achieving professional buy-in by teachers, curriculum specialists, and administrators. State and local effort has produced some truly outstanding exemplars as to how to set ambitious curricular goals and develop state-of-the-art techniques to measure them. Pittsburgh and Vermont are vivid instances. It would be a shame to lose this innovation; NAEP itself would suffer with the loss of examples and potential collaborators.

In trying to project a scenario here, much depends on *how* local NAEP would be achieved. If linking and equating were used to achieve local reporting of NAEP scores, state and local programs could continue to be distinctive, while tying into the NAEP system. If NAEP were packaged and extended homogeneously, distinct local and state programs could be driven out.

2. If NAEP is used as a high-stakes accountability device at the local level, this could lead to teaching to the test and distortion of the curriculum.

We have ample experience in the U.S. that tests applied in a high-stakes accountability environment can result in inappropriate responses by teachers and school administrators. Enough examples exist of using actual tests materials to teach, changing test scores, or focusing student work on learning how to take tests to be sure of this. Obviously, when these things happen, priorities are wrong, and instruction is diverted from more important and legitimate goals.

What are the answers to these fears? Two will be put forth here. One is that focusing instruction on the content of the tests is appropriate, if the tests contain or reflect what we want students to learn. The other is that we know how to develop and use tests in ways that would reduce misuse.

First, the corollary to the argument on national curriculum made above is that it is not necessarily bad to have tests and accountability mechanisms focus attention on desired knowledge and capabilities. Indeed, that is what we are trying to do with an accountability system. The key point is whether we agree on where we want the system to go and what we want students to learn. If we agree on a common set of instructional targets or goals, then tests and other vehicles can be used to strengthen attainment of the targets. We are at a point in this country where that kind of consensus exists, and accountability can be used, along with support for instructional improvement, to move the system in those directions.

How should tests be used to accomplish this? The main limitation to previous achievement tests, which resulted in their misuse, was the difference between the content of the tests and the content we wanted taught. Standardized, multiple-choice achievement tests have over-emphasized relatively low-level operations, which are amenable to short-answer test item formats. Scores based heavily on this kind of item and representing this kind of skill become measures of achievement in reading, mathematics, or other subjects. Then, when the pressure is put on, it is natural for educators to focus on those narrow aspects of the curriculum, on which they know the tests depend, to bring scores up.

Tests that would avoid these problems and misuses would be tests that fully represent the curriculum and that measure important knowledge and capability through means that are not specifically predictable. Stories are circulating about responses by teachers to New York's first hands-on science tests, where activities just like the test exercises were used in classrooms after the first test administrations. This practice would miss the larger goals tapped by the exercises, while failing to prepare students for the specific task that would be used in the future. Tests need to fully tap the domain to be measured and do it in ways that students and teachers understand to be fair but unpredictable in the specific form they will take.

Designing tests that fully represent a domain is another impetus behind innovations in assessment. Many topics or targets in science, reading, mathematics, social studies, writing, the arts, or foreign languages have been left off tests because they defied standard item formats, scaling, or other testing practices. It is to include these targets in testing that has prompted the incorporation of actual compositions in writing assessment, open-ended comprehension questions in reading, verbal problem-solving questions in mathematics, and laboratory exercises in science.

Another factor in precluding misuse of tests is taking care in how they are presented to teachers. Virtually no consideration has been given to how incentives to perform on standardized tests should be conveyed to teachers. Instead, accountability systems have applied the pressure and let local staff work out their responses. NAEP could avoid this by crafting messages to teachers that would head off inappropriate responses and encourage appropriate ones. One way to do this would be letting teachers know that the tests represent a domain and explaining what that domain is. Explaining that the test will be a random selection in terms of the specific items and exercises used is another.

3. Use of NAEP at the local level could interfere with the national program.

We have already discussed the issue of teaching to the test frameworks and specific tests. In at least one other regard, using NAEP at the local level could pose problems for the national program: in item security. NAEP uses a matrix sampling design with an evolving item pool, in which some new items are included each cycle. The design depends on security for the existing items. If NAEP were used for high-stakes, comparative testing at the district or school level, keeping these items secure would be difficult. Use in this way would probably require development of many more new items each cycle, so each assessment would be composed of relatively large shares of items that were new. How this would affect NAEP's ability to produce trend lines would have to be considered.

Another effect would be in the student and exercise sampling design. To represent schools or small districts in the NAEP grade levels, a different sampling and spiralling approach could be needed. How this would affect the scaling and analysis of the results for both the local and national (or state) program would have to be considered. It should be noted that in the 1990 state-by-state program, the District of Columbia and several sparsely populated states participated and were represented, so medium-to-large jurisdictions *could* be represented under the current design, which was about 2,000 students per grade level.

4. The cost of a local NAEP could be high.

If NAEP were administered at the local level, the cost of the whole NAEP program would be very large. NAEP generally costs more on a per-pupil basis than commercial, standardized tests. How costs could be split between local, state, and federal sources would have to be worked out. However, local use of NAEP would supplant much current state and local testing. If NAEP could serve for some of those layers of testing, savings could be achieved.

Overall Assessment of Impact

How can some of these benefits and potential liabilities be weighed in coming to a conclusion about whether NAEP should be used at the local level? Two of the factors to be weighed most carefully are the informational value of NAEP against its potential for misuse. The informational value offered by NAEP at the local level is very large. New forms of information--high-quality student achievement data and instructional program background data that are not now available--would be made available. If this happened on a wide scale, which states could help accomplish, local schools and districts would have this information in a powerfully comparable form.

On the debit side is the potential for misuse. We have little information of how prevalent teaching to tests and distortions of practice as a result of accountability pressures are now. Gross instances of cheating appear to be fairly rare, but many teachers report feeling pressed to obtain higher test results. Local use of NAEP would have to be implemented in thoughtful accountability systems, where consequences are realistically and carefully thought out. In addition, the testing program would have to continue farther in its current direction of becoming an assessment program that is difficult to teach to inappropriately. NAEP is as close to being this kind of testing program as any standardized program in the country right now.

The other liability is whether implementation of local NAEP would drive out good local and state testing programs. This depends heavily on whether development of local NAEP is done through linking and equating systems. If linking became available and economically feasible, it could be expected that states would use it to maintain their testing programs wiath their particular features and purposes, while tying into NAEP and generating NAEP scores for schools and districts. States would be more able to afford and to have the political will to do this than locals, but states determine the landscape in assessment now, anyway.

The incentives for states to maintain their own assessment programs should not be overstated. Three of the most interesting state assessment development programs in recent years--Vermont, Kentucky, and Maryland--emerged in large part because a local version of NAEP was *not* available. It would appear to be in the interest of assessment generally, the NAEP program, and states for states to be encouraged and enabled to develop statewide systems that are distinct and creative, while tying into a national assessment system which provides local schools and systems valuable data in a common national currency.

References

Burstein, L. et al. 1985. *Using State Test Data for National Indicators of Education Quality: A Feasibility Study*. Center for the Study of Evaluation, UCLA, Los Angeles.

Council of Chief State School Officers. 1986. *Summary: State Collection Practices on Universe Data Elements*. Education Data Improvement Project, CCSSO, Washington, DC.

Council of Chief State School Officers. 1988. *Assessing Mathematics in 1990 by the National Assessment of Educational Progress*. National Assessment Planning Project, CCSSO, Washington, DC.

Council of Chief State School Officers. 1990. *Reading Framework: 1992 National Assessment of Educational Progress Reading Assessment*. 1992 NAEP Reading Consensus Planning Project, CCSSO, Washington, DC.

International Reading Association. May 1991. Delegates' Assembly Resolution.

National Assessment Governing Board. 1990. *Writing Objectives for the 1992 NAEP Assessment*. Washington, DC.

National Council of Teachers of Mathematics. 1991. *Professional Standards for Teaching Mathematics*. Reston, VA.

Pecheone, R. 1986. Presentation to Connecticut State Conference on Testing.

Cautions on the Future of NAEP: Arguments Against Using NAEP Tests and Data Reporting Below the State Level

Walter Haney
George F. Madaus
Center for the Study of Testing, Evaluation and Educational Policy
Boston College

December 1991

I. Introduction

For more than two decades the National Assessment of Educational Progress (NAEP) has served as our nation's preeminent source of information on the learning of students generally, and our most accurate source of national data on changes in achievement levels over time. Over the past decade NAEP's capability to provide accurate information to policymakers and the public stands in sharp contrast to many state and district high-stakes testing programs implemented during the late 1970s and 1980s in response to a growing dissatisfaction with our schools. These programs were generally designed to use test results as means of school accountability or as bureaucratic mechanisms to make important decisions about individual students (e.g., to track, retain, place, and certify), teachers (e.g., to evaluate teachers or award merit pay), schools or school systems (e.g. to allocate funds or other resources, rank schools, place schools or systems into educational receivership). It is now widely recognized that such high-stakes state and district testing programs have had serious unintended, but predictable, negative outcomes on students, teachers, schools, and school systems (American Educational Research Association [AERA] 1991; Darling-Hammond 1991; Jaeger 1991; Koretz, Linn, Dunbar, and Shepard 1991; Madaus 1991; Shepard 1991; Stake 1991). Moreover, these programs, unlike NAEP, have in at least some instances clearly yielded inaccurate and misleading information about how well our schools and students were performing.

At present, however, policies and attitudes toward NAEP's role in American education are changing in ways that we believe may endanger the status of NAEP as the premier indicator of our nation's educational achievements and progress. Such changes are apparent in the following three developments:

- **Trial State Assessment**. NAEP has been expanded to permit state-by-state comparisons; currently 44 states are participating in this effort. While this effort is still at an informational or descriptive level, the recent ranking of states on the basis of state-level NAEP data, widely reported in popular newspapers and magazines, could easily energize national or state-level policymakers to use the results to allocate rewards or sanctions for performance.[1]

- **Performance standards**. The National Assessment Governing Board (NAGB) has set performance standards for each of three grade levels (4, 8, and 12) in mathematics. This modification is designed to allow NAEP to report how well students are doing compared to agreed-on national standards of excellence (O'Neil 1991). These standards could easily be used to trigger important administrative decisions at the state level. Further, such standards could be the trigger for critical school choice and receivership decisions if NAEP were extended downward to district or school levels, or for placement, retention, and certification decisions if NAEP were further extended downward to the individual student level.

- **Development of subject frameworks**. Subject frameworks are being developed which could form the basis for a national--and, as we will argue, in effect, federal--curriculum.

The National Academy of Education (NAE) is now undertaking scrutiny of current and proposed changes in NAEP. As part of that undertaking we have been asked by the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment Project to prepare a position paper that "argues the case against lifting the prohibition on reporting [NAEP] data below the state level" (Glaser and Linn—personal communication, July 12, 1991). There are two current proposals to lift the present prohibition against the use of NAEP below the state level. The first is a recommendation by the National Assessment Governing Board (NAGB). The second, a more radical proposal for changing NAEP, is contained in the *AMERICA 2000* report issued by the U.S. Department of Education. These proposals are the focus of our analysis.

Our analysis is organized as follows. It opens with a summary of the two proposals: the NAGB proposal to lift the prohibition against using NAEP below the state level, and the *AMERICA 2000* proposal for an individual version of NAEP while the

---

[1] We should note that we have not yet had the opportunity to review the technical report on the Trial State Assessment in detail, but we will draw on examples from the Trial State Assessment later in this paper.

American Achievement Tests are being developed. (From this point on we shall use the simple shorthand phrase "lifting the prohibition" in place of the longer phrase "lifting the prohibition against the use of NAEP tests and data reporting below the state level.") Second, we review the rationale behind NAEP's original design, which was explicitly planned to prevent use of NAEP to judge states, schools, or individuals. Third, we recount some of the practical problems that likely would arise if the prohibition were lifted. Fourth, because we feel that both the NAGB and *AMERICA 2000* proposals would transform NAEP into an extremely high-stakes testing program--a key premise in our analysis--we summarize how high-stakes testing programs play out in terms of six general principles. Finally, we summarize our position on prospects for reporting NAEP results below the state level and sketch out our own view of a more fruitful way of thinking about how any investment in educational testing and assessment might be considered with regard to its potential for improving teaching and learning in the schools of our nation.

Before we set out on this itinerary, two additional explanations are needed. First, we do not attempt to describe the current nature of NAEP or to document what seems to us to be the clear standing of NAEP among specialists in educational testing and measurement as the single best source of national data on the learning of our nation's students. Instead we simply presume that our readers have a reasonably good understanding of the current structure of NAEP and that they concur with our assessment about the high opinion in which NAEP is currently held.[2]

Second, we make a distinction between the NAEP program itself, and some of the testing, sampling, and reporting technology developed as part of that program. Despite our reservations about prospects for extending NAEP downward, as called for by NAGB and in *AMERICA 2000*, we think that the technology developed in NAEP might very usefully be applied in other kinds of testing programs in the United States. Among technical innovations developed in NAEP that might well be adapted for other purposes or assessments are the consensus process for determining the assessment objectives, open-ended and essay items, portfolio measures, student and item sampling technology, item response theory scaling and subscaling, estimation by plausible values, developmental scales, scale anchoring, error estimation, and the reporting system.

---

[2] One development related to NAEP should be noted. Although even some people directly involved in NAEP are unaware of it, it appears that an individual version of a NAEP instrument has already been developed and is being marketed commercially. We refer to the Tests of Applied Literacy Skills (TALS), which grew out of the 1985 assessment of adult literacy. While the new TALS tests seem to represent a useful addition to the armamentarium of instruments available for assessing adult literacy, they also raise a number of questions that ought to be addressed before proceeding with development of any more individual versions of NAEP tests. However, since the TALS tests are intended for out-of-school adults, rather than school children, we will not elaborate further on the development of TALS in this paper.

II. Proposals to Remove the Prohibition on Reporting NAEP below the State Level

Does it make sense to report NAEP results below the level of the the nation or states? Addressing that question requires first that one consider how NAEP results would be used at the district, school, or individual levels. That is, what descriptions, decisions, and consequences for a district, a school, and individuals will flow from the results? Some may believe that tests, like mountains, are simply there to be scaled, reported and used for whatever purposes one wishes. But it seems to us that when it comes to educational policymaking generally, and educational measurement and testing specifically, the anarchist philosophy of the mountain climber is not reasonable. Thus before we can address the question of whether it makes sense to lift the prohibition we must consider questions of use and consequences that will most likely ensue.

The NAGB Proposal

On November 17, 1990, the National Assessment Governing Board (NAGB) reaffirmed its December 9, 1989, position that Congress should remove the prohibition against use of NAEP tests and data reporting below the state level (National Assessment Governing Board 1989, 1990). NAGB proposes that a local option be given to states, districts and localities to use NAEP and specifically that the 1994-1996 "NAEP tests and items should be available to the states, districts, schools or organizations of schools that contract and pay for them. Publication of individual student scores should continue to be prohibited" (NAGB 1990, 2).

While NAGB supports use of NAEP by states, districts, and schools, it does not spell out how the results would be used. However, NAGB does note that it is at the district and school building levels where many important decisions are made, and that NAEP data would provide helpful information in informing those decisions. In addition, NAGB suggests that NAEP can provide valuable information to parents and others interested in school improvement (NAGB 1989). Thus, it appears that NAGB would, initially at least, leave the use question up to the state education agencies (SEA) and local education agencies (LEA).

NAGB's position on the use of NAEP results with individuals is somewhat confusing. Although NAGB affirms the prohibition against publishing individual scores, it goes on to discuss how differences in motivation and possible preparation for NAEP by LEAs would present a serious problem when comparing the results of students who took NAEP knowing they would get individual scores with students in the national and state samples who are assured NAEP cannot "count" for them (NAGB 1990). This could be read to imply that while individual scores cannot be published, they can be given to individuals and that the SEAs or LEAs that opt into the NAEP program could use the results to make decisions about individuals as long as the individual scores are not published.

It is unfortunate that NAGB does not confront and discuss the myriad of ways SEAs and LEAs might use district, school, and individual NAEP results. Based on past experience with state and local testing programs, it would not be surprising to see SEAs and LEAs attach the following high-stakes consequences to NAEP results:

- putting districts or schools into receivership,
- making interdistrict and interschool comparisons,
- using results in a school choice plan,
- allocating resources on the basis of results,
- holding teachers accountable, and
- making important classification and certification decisions about individual students.

Based on past form, given the freedom to use NAEP as they saw fit, many SEAs and LEAs would likely use the results for high-stakes decisions about districts, schools, or individuals.

AMERICA 2000 Proposal

The most straightforward declaration of how NAEP would be used below the state level is found in *AMERICA 2000* as part of its larger strategy for national educational reform. The *AMERICA 2000* strategy is comprised of four parts, most of which we cannot comment on here. We note, however, that we concur with some of the *AMERICA 2000* proposals.

The segment of the *AMERICA 2000* proposal which is directly relevant to the future of NAEP, and with which we sharply disagree, is related to the call for new "American Achievement Tests" (AAT). The AAT are meant to cover "core subjects" of English, mathematics, science, history, and geography and to be based on "new world standards." President Bush has asked that the first tests be available by September 1993 (Miller 1991). *AMERICA 2000* clearly spells out the proposed uses of the AAT:

- "to foster good teaching and learning"
- "to monitor student progress"
- to help determine college admissions
- to aid employers in hiring decisions
- to check the nation's progress towards the National Educational Goals developed by President Bush and the National Governors Association (U.S. Department of Education 1991, 11)

One may ask, "What does the AAT have to do with NAEP?" *AMERICA 2000* reasonably notes that since the AATs probably cannot be developed by the target date of 1993, the Department of Education should ask Congress to authorize the rapid deployment of individual versions of tests used by NAEP, at least in reading, writing, and mathematics. Thus, in effect, the *AMERICA 2000* plan proposes that, at least in the short term, an individual version of NAEP should be used to promote good teaching and learning, to hold schools and school systems accountable, to determine college admissions, to influence employment decisions, and to check educational progress at the national level.[3]

The idea of using any test, be it an individual version of NAEP or any other type of test, for such a range of purposes is severely ill-founded. It is rather like using only one tool, for example, a hammer, for jobs as different as brain surgery and pile driving. We currently have different tests for evaluating educational programs, for college admissions and for employment hiring, for the very simple reason that tests for such different purposes need very different characteristics. This is one of the main reasons why the *Standards for Educational and Psychological Testing* (AERA 1985) of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education prescribe that tests cannot be validated in the abstract. Instead, tests must be validated for their intended uses. Similarly the Uniform Guidelines on Employment Procedures (1985), promulgated by the Equal Employment Opportunity Commission, the U.S. Department of Labor, and the Office of Personnel Management of the Executive branch of the federal government (and recognized in numerous U.S. Supreme Court decisions), prescribe that employers must validate their employment tests in terms of their relevance to the specific job for which people are being selected. Employers asked to use an AAT version of NAEP to make hiring decisions could rightly ask why, since for almost three decades the federal government has told them that they must validate employment tests for the particular jobs to be filled.

---

[3] Since the new tests are to be used to monitor progress toward the President and governors' national education goals, then the NAEP tests would of necessity be geared to grades 4, 8, and 12. Those are the specific grade levels at which the national goals called for "demonstrated competency in challenging subject matter including English, mathematics, science, history, and geography" (U.S. Department of Education 1991, 9). In light of the fact that NAEP has historically been aimed at assessing the learning of 9-, 13-, and 17-year-olds (or more recently 4th, 8th and 12th graders), it is probably not a coincidence that NAEP has been proposed as the initial manifestation of the new American Achievement Tests.

Putting aside professional standards about test validity and federal regulations regarding employment testing (which, after all, could be changed), simple common sense tells us that a single test--an individual version of NAEP or something else--cannot serve equally well all of the uses proposed in *AMERICA 2000* or the many uses that SEAs and LEAs could mandate under the NAGB plan to lift the prohibition. To illustrate this point let us simply sketch out some of the characteristics needed in assessments aimed at serving different purposes.

Since space does not permit us to discuss the whole range of purposes intended to be served by different kinds of tests, we focus on just three purposes that are often ascribed to tests and assessments. These three are set out directly in the *AMERICA 2000* proposal for the the AAT (and, by extension, the individual version of NAEP) and by inference in the NAGB proposal. These functions are:

- providing information for the purpose of accounting to people outside of schools,
- providing information to teachers to help them improve instruction, and
- helping students learn.

Providing information for purpose of accounting to people outside of schools. School accountability is one of the most frequently discussed purposes of testing. Indeed in *AMERICA 2000*, the proposal for the AAT is included as part of a "15-point accountability package" through which "parents, schools and communities can all be encouraged to measure results, compare results and insist on change when the results aren't good enough" (U.S. Department of Education 1991, 11). Similarly, when NAGB asks that NAEP results be used below the state level, it specifically asks the Congress to repeal subsection (4)(C) of P. L. 100-297 which prohibits using results to "rank, compare, or otherwise evaluate individual students, schools or school districts" (NAGB 1989). Ranking, comparing, and evaluating are, of course, all accountability functions.

Accountability is one of the least important functions for testing and assessment, especially when the overriding purpose is to improve instruction and learning. There are two broad reasons for this general conclusion. First, schools are already held accountable in myriad ways to many audiences. Schools account to parents about their children's performance via report cards and parent teacher meetings. Schools must account to school boards that are elected or appointed to oversee them. They also are asked to account to state officials (for example, regarding student attendance) and to comply with numerous state regulations (for example, regarding the length and number of school days). Moreover, schools must comply with various federal regulations and if they receive federal money via categorical programs (such as Chapter 1), report on the efficacy of their expenditures. Most schools also are subjected to scrutiny, at least irregularly, by the popular press and by numerous interest groups, ranging from parent groups to groups for and against religion, the theory of evolution, sex education, or cultural pluralism.

In short, despite much rhetoric lately about the need for more school accountability, schools are already held more accountable and in more ways than are most other social institutions. Indeed, the *AMERICA 2000* report implicitly acknowledges that schools are overly constrained by too many requirements and regulations when it points out that "The individual school is education's key action-and-accountability unit. The surest way to reform education is to give schools and their leaders *the freedom and authority to make important decisions about what happens* . . . " (U.S. Department of Education 1991, 27, italics added). High-stakes uses of NAEP, which are directly proposed in *AMERICA 2000* and which can be inferred in the NAGB's proposal, can deny educators their freedom to make many important decisions. In fact, high-stakes uses of NAEP are proposed because policy makers do not trust educators to make many important decisions about districts, schools, teachers, and pupils.

Nonetheless, many proponents of greater school accountability through testing argue that in the past the wrong standards--procedural requirements, financial regulations, attendance, and the provision of certain input variables--were used to hold schools, educators, and pupils accountable. Now, proponents say, schools must be held accountable for student performance.[4]

The problem with test-based accountability is that test results cover only a small part of the range of results for which most people think schools should be held accountable. President Bush and the governors acknowledge this fact in the national education goals, established in 1990 and out of which the *AMERICA 2000* strategy evolved. Two of the six goals pertain to student learning: goal 3 says that students should have demonstrated competence in challenging subject matter; goal 4 says that U.S. students should be first in the world in science and math. However, two of the other goals clearly suggest other standards by which schools should be held to account:

> Goal 2: By the year 2000, the high school graduation rate will increase to at least 90 percent.
>
> Goal 6: By the year 2000, every school in America will be free of drugs and violence and will offer a disciplined environment conducive to learning. (National Education Goals Panel 1991)

Public opinion polls indicate that when it comes to assessing school quality, school characteristics other than test scores are what concern people first and foremost. For example, when the latest Gallup poll queried a national sample of people about attitudes toward the six national educational goals, highest priority was given to the last goal: to free every school from drugs and violence and offer a disciplined learning environment (Elam 1990). Similarly, when asked what aspects of a school would be most influential when deciding on a school for their children (should parental choice of school be adopted in their community), the respondents gave substantially higher priority to quality of the teaching staff, maintenance of school discipline, curriculum, and size of classes (rated as "very important" by 87, 78, 73 and 56 percent of the national sample) than to "grades or test scores of the student body" (which was rated as "very important" by 48 percent of the sample; Elam 1990, 44).[5]

---

[4] In his new book, *We Must Take Charge: Schools and Our Future*, Chester Finn illustrates this view of accountability when he argues that holding schools accountable for outcomes "is the only kind of accountability worth having in 1991" (Finn 1991, 149).

Thus, the second reason for thinking that holding schools accountable in terms of test scores should not be the overriding function of assessment is that there are other important standards by which people judge schools. Students' test scores (or assessment results) simply do not illuminate these other standards. Indeed some research indicates that testing programs may work against some of these other goals. Some research, for example, indicates that student competency testing and graduation testing programs may spur dropouts from school (Catterall 1989; Kreitzer, Haney, and Madaus 1989).

Moreover, considerable research suggests several reasons why group average test scores may present a very inaccurate picture of the extent to which schools or school systems are educating students. One is the simple fact that group averages on tests tend to be higher when larger proportions of students are excluded from school or from testing. Thus, the states with the highest college admissions test scores tend to be those in which smaller proportions of students take college admissions tests (Powell and Steelman 1984). Similarly, countries which tend to rank higher in international test comparisons tend to be those that educate smaller proportions of their students at higher grade levels (Rotberg 1990, 1991).

A second limitation with group average test scores as indicators of school quality is that simple comparisons of school average test scores can be highly misleading. Not only are such averages related to the socioeconomic backgrounds of students attending different schools, but they are also affected by the level of learning of the schools' incoming students, the extent to which teaching matches the content of the test, and schools' grade retention policies. One recent review of such complications in comparing schools on the basis of average test scores offered the following advice:

---

[5] Even though he cites evidence from Gallup polls in more than a half dozen places, Finn (1991) seems to have overlooked the Gallup poll evidence summarized in the paragraph above when he writes: "Most Americans are accustomed to gauging the performance of their school systems by whether those [test] scores are rising or falling" (99). Though Finn seems to wish otherwise, test scores do not seem to be the primary criterion by which people judge school quality.

> . . the only valid way in which assessment systems can be used to compare schools, and make inferences about a school's effectiveness, is to use longitudinal achievement data, with other variables known to affect achievement adequately represented in the analysis, and with student outcome measures that are systemically valid. Much hard work remains to be done before these goals can be realized. Meanwhile . . . assessment systems should not be reporting school comparisons. That may do more harm than good (Cooley and Bernauer 1991, 168-9).[6]

Impatient education reformers doubtless view such cautions as minor and temporary technicalities or dismiss them as academic nitpicking. But even if we put aside as a minor technicality the fact that school average test scores have not been shown to be valid indices of schools' effectiveness, it is worth inquiring further into the characteristics of assessments needed for the purposes of school accountability. First, for purposes of school accountability, people outside the schools generally want only summary information. They want to know how schools are doing in general, and not every day or week, but maybe annually or semi-annually. Corporations, after all, account to their shareholders only via annual reports and annual meetings and typically report to financial analysts with only quarterly reports on business. Similarly, school outsiders likely do not want detailed information on how each and every student is scoring in each subject studied, but instead desire information on how students are performing in basic academic areas, such as English and math. This is one of the prime reasons why standardized tests such as college admissions tests and achievement batteries such as the California, Iowa, and Metropolitan achievement tests have never been keyed to the curricula of particular schools, but instead to the broad and commonly studied subject areas such as reading comprehension and math computations. Thus, for purposes of accountability, people generally want to know not just how a school is performing in terms of that one school's particular curriculum, but how it is doing in comparison with other similar schools in broad curriculum areas such as English and math.

Providing information to teachers to help them improve instruction. Even brief reflection indicates that assessment results that would be useful to teachers in helping them to improve instruction need very different characteristics than those sketched out above. Before discussing this point, however, let us briefly summarize evidence on the question of the extent to which teachers in the past have found standardized test results to be useful.

---

[6] For a similar critique of the value of school average college admissions test scores as indicators of school quality, see Fetler (1991).

Since mid-century there have been a number of studies of how tests are actually used within schools. (See Haney 1984 and Madaus and Kellaghan 1990 for summaries of much of this literature.) Several themes emerge from this literature. First, teachers (and school administrators) are typically not terribly knowledgeable about the technology of testing, with misinterpretations of reported scores (such as grade equivalents) quite common. Second, run-of-the-mill testing programs with no clear consequences attached to them are typically seen as not very useful. Salmon-Cox (1981) reported that teachers do not find test results to be of much use because "Teachers desire diagnostic tests that are precise, closely matched to curricula and instruction, and timely. Achievement tests of the kind now widely used do not match these criteria" (634). In addition, a variety of research indicates that standardized test results do not tell teachers anything they did not already know about their students because teachers' ratings of students, independent of test scores, correspond substantially with how students perform on standardized tests (Archer 1979; Hopkins, George and Williams 1985; Kellaghan, Madaus and Airasian 1982; Pedulla, Airasian, and Madaus 1980).

There are two broad circumstances under which teachers' instructional behavior is affected by students' test results. The first is before a teacher has personal acquaintance with a student (Raudenbush 1984). When a teacher is just beginning to teach a particular student and the teacher has few if any other sources of information about the student, knowledge of students' standardized test results may affect the teacher's teaching.

The other circumstance in which teachers' instruction is affected by test results is when large consequences, for students, teachers, or schools, are attached to test results--or what has come to be called high-stakes testing. Under this condition, teachers will tend to focus instruction on the content and form of the important test (Hastings, Runkel, Damrin, Kane, and Larsen 1960; Madaus 1988; Popham, Cruse, Rankin, Sandifer, and Williams 1985; Rafferty 1985). Indeed, if the stakes attached to test results are large enough, history shows that some teachers and administrators will outright cheat on tests (Cannell 1989;

Radwin 1981).[7] We discuss effects of attaching high stakes to test results in a subsequent section of this report.

Helping Students Learn. This is a function which is often claimed for tests but is in fact rarely considered in detail. If there is a theory implicit in *AMERICA 2000* as to how the AAT will help students learn, it appears to be that by attaching real rewards for passing and real consequences for failing the AAT, students will be motivated to learn more. This argument is strikingly similar to the arguments of the utilitarians that surrounded the introduction of examinations in 19th century Britain: self-interest is the main motive for study, and since study involves drudgery it is necessary to link important rewards or sanctions to successful test performance (Madaus and Kellaghan, in press). Thus, fear of a low score on a certification test, and the negative implications for further education or employment associated with such a score, will motivate a target population of lazy, recalcitrant or otherwise unmotivated students to work hard (Educate America Inc. 1991; Finn 1991; Resnick 1991; Shanker 1990 ; Smith, O'Day, and Cohen 1990). Finn puts it this way, "When it comes to academic learning, I believe that external consequences are the main determinant of how hard [students] work" (Finn 1991, 125).

The flip side of the motivational argument however, is that positive motivational effects are likely to operate only if students believe that they have a good chance of achieving the rewards attached to high test performance. For students who are not likely to do well, the possible negative effects are of serious concern. For example, there is evidence that competency tests for graduation, or test-based retention policies, are implicated in increasing the drop-out rate (implying negative motivation) (Haney and Madaus 1989; Kreitzer, Haney, and Madaus 1989; Madaus and Greanery 1985; Metropolitan District Commission [MDC] 1988; Shepard 1991). In addition, a survey of students in grades 2 through 11 in four large states found "negative impact [of testing] on students that can be summarized in three general trends: growing disillusionment about tests, decreasing motivation to give genuine effort [on tests], and increasing use of inappropriate strategies (Paris, Lawton, Turner and Roth 1991, 14). Moreover in a second survey in another state, the same investigators found that

---

[7] It is worth noting that despite much publicity recently to instances of teachers cheating on tests (including one "60 Minutes" television program devoted to a South Carolina teacher who got fired for cheating), the vast majority of educational personnel responding to an anonymous national survey indicated that less than 10 percent of teachers in their school districts engage in "non-standard" testing practices (Hall & Kleine 1989).

> the results of standardized tests become increasingly less valid for low achievers, exactly the group who are most at risk for educational problems and who most need diagnostic testing. Their scores may be contaminated by inappropriate motivation and learning strategies that further debilitate their performance and affirm a self-fulfilling prophecy of low scores. Apparently in their efforts to decrease personal anxiety and increase the protection of their own self-esteem, they relinquish effort and appropriate strategies on standardized achievement tests (16).

NAEP would be used at the fourth grade level in the *AMERICA 2000* proposal; SEAs and LEAs could attach high-stakes to fourth grade performance under the NAGB proposal to lift the prohibition. The positive power of high-stakes tests to motivate high school students cannot be uncritically extended downward to fourth graders. There is very little empirical evidence about the motivational potential of tests with important but perhaps rather abstract rewards or sanctions on *young* children from very diverse backgrounds.

The problem with the motivational theory of high-stakes assessments as a strategy for school reform is that, apart from anecdotal evidence, there appears to be little evidence that lack of motivation to do well in school is a key problem in students' learning. One major study of motivation and student effort in a major urban school district, for example, concluded:

> Students in our studies reported that evaluations of their performance in school were central to their life interests. . . . Evaluations of performance were seen as affecting important material sanctions in the future, in the form of jobs and careers, as well as important social sanctions in the present in the form of the opinions of parents, counselors, and friends. Moreover these sanctions were considered as very important by most students. . . .
>
> Low achieving students, as well as minority group students, were just as likely as high-achieving students and Anglo students to report that evaluations received in school were influential. Since most students in our studies perceived evaluations of their school performance as being influential, low student effort and low achievement in school cannot be attributed to lack of influential evaluations. (Natriello and Dornbush 1984, 137-8)

Thus it appears that the "theory" behind how new national tests in general and a district-, school-, or individual-level NAEP in particular will increase student learning via higher stakes attached to test result is based on a faulty premise. If low student achievement in school cannot be attributed to lack of influential evaluations currently, it is hard to envision how more influential evaluations via national tests and external sanctions could possibly improve matters.

If high-stakes testing does not improve student learning indirectly via motivation, it is worth considering how tests might aid student learning more directly. Though much standardized testing in schools is often described as "educational," the features that would facilitate student learning are notably lacking. This is clear simply from a consideration of some of the essential elements needed to facilitate learning. To acquire new knowledge or master a new skill, students need both opportunities for practice and rapid, detailed feedback on how they are doing.[8]

Conditions conducive to learning have been recognized for many decades in many theories of learning, for instance in Edward Thorndike's (1913) law of effect, which holds essentially that learning is enhanced when people see effects from what they try. Research has also identified conditions promoting learning. Malone (1980) studied what makes for intrinsically motivating educational computer games and identified motivating attributes such as rapid feedback, variable difficulty providing appropriate challenge, and randomness in presentation. A meta-analysis of 40 previous studies on the instructional effects of feedback in test-like events showed that feedback provided rapidly (e.g., immediately after a test was completed) was more effective than feedback delayed a day or more. Also, feedback that identifies or explains correct answers is more instr[illegible]nally effective than when learners are simply told whether their answers were right or wrong. (Bangert-Drowns, Kulik, Kulik, and Morgan 1991)

All of this--common sense, theories of learning, research on intrinsic motivation, and studies of instructional feedback--clearly indicate that if the NAEP prohibition were lifted, the resultant tests would not contribute directly to student learning. Students would not get rapid or specific feedback on their work, but only summary scores, and only after months of delay.

---

[8] Of course, these requirements are not something that *necessarily* needs to be supplied by an external authority such as a teacher (much less testing company or scoring service). Indeed, surely one aim of education and child-rearing generally is that people should become autonomous and self-regulating adults, capable of judging the success of their own endeavors, rather than attending forever to external sanctions and the judgments of foreign authorities.

Our critique of the *AMERICA 2000* and NAGB proposals for versions of NAEP to be reported below the level of states boils down to the simple logic that that no test can reasonably serve the range of purposes suggested or implied. Not only does such a proposal violate standards set out in the *Standards on Educational and Psychological Testing* (AERA 1985), and the *Uniform Guidelines on Employment Procedures* (Equal Employment Opportunity Commission, U.S. Department of Labor, and Office of Personnel Management 1985), but even the simple logical requirements of tests aimed at the specific purposes of holding schools accountable, helping teachers teacher better, and helping students learning. As a means of highlighting this point, Table 1 lists the differential features needed in tests aimed at these three purposes.

Table 1. Features needed in tests for three different purposes

	PURPOSE		
FEATURE	School accountability	Improving teaching	Promoting individual learning
Degree of detail	Summary, not detailed, information	Detailed feedback on individuals; patterns among small groups within class	Detailed and specific to problems tried
Frequency/ Feedback	Relatively infrequent (1-2x/yr)	Rapid feedback within hours or days	Immediate, after test or problems tried More than score or right/wrong feedback
Basis	Not specific to a particular school's curriculum	Based on course, diagnostic with regard to what to teach or do next	Problems tried, specific misconceptions

III. Reasons behind the Original Reporting Prohibitions on NAEP

A detailed account of the early history of NAEP is beyond the scope of this paper.[9] However, we will briefly highlight some of the early history and arguments bearing on the founding of NAEP and more recent evidence bearing on such arguments.

The 1867 Act establishing the Department of Education recognized the need for gathering descriptive information about "the condition and progress of education in the several states and territories." Of course, at that time testing as we now know it did not exist. From the 1920s to the 1960s, standardized tests had little or nothing to do with state or federal policy. It was not until the early 1960s with the passage of the Elementary and Secondary Education Act of 1965 and the establishment of the National Assessment of Educational Progress (NAEP), that the Department began systematically to gather test data as part of its original mandate. Further, state departments of education have only recently begun systematically to collect test data to describe the status of education.

Though it is often forgotten, when NAEP was first proposed in the early 1960s, its advocates initially envisioned state and even district level descriptions of performance. For example, in the very first conference leading to the founding of NAEP, in December 1963, the question arose as to the level at which national assessment results might usefully be reported. In response, Francis Keppel, then U.S. Commissioner of Education, and Ralph Tyler, the person most responsible for the founding of NAEP, commented that "We will have to end up obviously with a state or school district or college breakdown" (cited in Hazlett 1974, 37).

Critics of early proposals for NAEP, such as Harold Hand, were quick to charge that whoever decided what went into the exercises would determine what would be taught in the schools. For example Hand suggested that early plans for NAEP were a "fiendish innovation" of the U.S. Office of Education and the Carnegie Corporation which "would lead to a national curriculum" (cited in Hazlett 1974, 95).

Although to the best of our knowledge early critics of NAEP did not make an historical argument, the history of public examinations certainly supported their position.

---

[9] Key sources regarding this history are Hazlett 1974 and Greenbaum, Garet, and Solomon 1977, plus other original source materials, some of which we will cite shortly.

There was then and is today considerable evidence across time and nations that an examination with important rewards or sanctions associated with performance eventually strongly influences what is taught and how it is taught (Madaus 1991; Madaus and Kellaghan, in press). One of many possible examples of this phenomenon is the French *Baccalaureat*. When first implemented the *Baccalaureat* program consisted of a catalog of 500 questions from which a number would be randomly selected each year to form the exam. From its earliest days therefore, the exam came to define what student should learn, and the catalog of questions eventually evolved into a highly centralized formal curriculum (Prost 1968).

In an article entitled "National Assessment Viewed as the Camel's Nose," Hand offered the following arguments against a national test: it would set obstacles to the realization of the goal of equality of educational opportunity; it would be the nose under the tent which would be followed by a monstrous camel in the form of a centrally controlled, dictated curriculum; it would stultify the curriculum; it would stultify local curriculum innovation and experimentation; and it would encourage cheating on the part of students and teachers alike (Beymer 1966; Hand 1965a: Hand 1965b; Hand 1966; and Saylor 1970).

During its first years of gestation, NAEP was viewed, if not with opposition, then at least with considerable suspicion by major education interest groups, including the American Association of School Administrators (AASA), the Association for Supervision and Curriculum Development (ASCD), the National Education Association (NEA), and the Council of Chief State School Officers (CCSO). Though many specific arguments were offered in opposition to NAEP, Hazlett recounts that opposition seemed based at root on the the following general argument:

> [National] Assessment was a *national testing program*, which would lead to establishment of a national curriculum which would postulate *national standards* leading to *conformity* of instructional objectives and learning, thus placing the *federal government in control* of education (Hazlett 1974, 380, italics in original).

Professional groups opposed to NAEP were worried that the national (and federal) influence of the new assessment would undermine professional judgments in both the determination of educational programs and also the tailoring of instructional objectives and curricula to meet individual students' needs (379-385).

In order to fend off early critics, NAEP's designers adopted several strategies. Tyler and others argued that assessment was fundamentally different than testing. In order to "purify" federal funding, NAEP was administered not by a federal agency but by the Education Commission of the States. Finally and most important for the purposes of this paper, procedural and technical safeguards were mandated that prohibited any individual, school, district, or state from being singled out for description or comparison (Tyler 1965). When the fears of the professional organizations that "Assessment was 'nationalizing' education and leading to federal domination were assuaged, they ceased their agitation" against NAEP (Hazlett 1974, 398).

The current experiment with state comparisons, and proposals to assess districts, schools, and individual students, alter the original safeguards that made NAEP possible and raise once again the arguments of NAEP's original critics that NAEP would be a national or federal test. If that happens and if important rewards or sanctions are associated with performance, then teachers will teach to the NAEP exercises and subject frameworks, which in turn will become the de facto federal curriculum of the nation.

Hand's prophecy of a central, though not federal, curriculum being imposed by a national testing program carried apocalyptic overtones. Today the idea is increasingly viewed by some educationists not only as acceptable, but a necessary development if educational reform efforts are to succeed (Finn 1991; Shanker 1990; Shanker 1991; Smith, O'Day, and Cohen 1990.) A curriculum or a test can be truly national but not necessarily federal, and this distinction is important to present advocates of a "national" curriculum (Finn 1991; Smith, O'Day, and Cohen 1990). For example, the Scholastic Aptitude Test (SAT) of the College Entrance Examination Board (CEEB), albeit private, is a national but not federal testing program, and several current proposals for a national exam or exam system envision a national, rather than a federal, testing program. (See Madaus 1991 for a description of other proposals for a national exam or exam system.) Further, one could argue that for some students the Advanced Placement Tests of the College Board define a national but not federal curriculum or syllabus of study.

NAEP is an interesting hybrid between a national and a federal test. NAEP was initially funded by the private foundations (mainly the Carnegie Corporation) but because of its huge cost it quickly became a program entirely funded by the federal government (Hazlett 1974). Sovereignty over the purse strings is the ultimate form of programmatic control. Since the federal government foots the bill for NAEP, NAEP is by the standard of purse strings clearly a federal examination program, and its standards and subject frameworks developed under federal monies are indeed federal.[10]

A federal test aimed at districts, schools, or individual students under the auspices of NAEP raises the central question that has not yet been answered and worse, has not even been asked by many proponents of new national tests. Exactly what will the results be used for? Will the results be used to certify, classify, sort, or select students--to make important decisions about them? How "voluntary" will an individual NAEP be and for how long? If NAEP is extended to the district, school, and individual level, and results are used for anything like the range of purposes suggested in *AMERICA 2000* or that might be possible under the NAGB proposal, the nation will soon have, for the first time, a truly national curriculum, and one which, it could be argued, could be called a federal curriculum, since it will be defined by the federally funded NAEP.

We note too that our concerns about the likely ill effects of a national high-stakes test and the lifting of the NAEP prohibition appear to be shared by many of our professional colleagues. At a June 1991 American Educational Research Association (AERA) Invitational Conference, research reports directed at the issue of a national test presented overwhelming evidence that in present high-stakes state-level testing programs the harmful aspects of such programs far outweigh the benefits. The participants at that meeting argued that while testing should be a major piece of systemic reform, the move to the use of high-stakes testing at the national level may be dysfunctional (See American Educational Research Association 1991); Darling-Hammond 1991; Jaeger 1991; Madaus 1991; Shepard 1991; Stake 1991). The research presented at the conference led the AERA Council to adopt the following resolution:

> The American Educational Research Association has concluded that a preponderance of research evidence suggests that America slow down, and think differently, about test-based educational accountability. (American Educational Research Association 1991)

Consequences of other high-stakes testing program help us predict possible social and individual consequences, and the technical and cost implications of a high-stakes national test. Misleading missionaries who make rosy predictions and bullish statements about any new technology generally underestimate the complexity of the problems associated with implementation, over-estimate the near-term promise, and under-estimate the impact and social consequences that follows (Brody 1991). This is currently what is happening regarding the proposal to lift the NAEP prohibitions. But it need not happen. We know much about the effects of tests on students. We have a wealth of evidence from two centuries of educational testing in Europe, the United States, and elsewhere. This knowledge comes from the testimony of teachers, administrators, students, and other observers to investigative panels and commissions. It can be found in the analyses of school inspectors; it comes from surveys and interviews of educators and students; it is embedded in a large corpus of popular literature--fiction, biography, poetry, drama, even the lowly detective novel--which describes how tests and examinations are experienced by teachers and students; and it can be found in empirical studies of the effects of testing. Before American students, already the most heavily tested in the world, are subjected to yet another testing treatment, we must carefully evaluate and apply what we already know about the effects of testing (Madaus 1991). A national test or testing system--including the one that would likely result if the NAEP prohibition were lifted--cannot be judged neutral a priori and must be scrutinized to see the ways in which conditions of power, authority, freedom, and social justice are deeply embedded in its technical structure, infrastructure, and implementation (Winner 1986).

---

[10] We are aware that individual states are bearing some of the costs of NAEP in the Trial State Assessment, but though we know nothing of the details of such funding arrangements, we suspect that such funding at least as yet has had little influence on the content of NAEP.

The erosion of the original safeguards, so vital to NAEP's birth, carries a clear lesson regarding the future of any national testing system. It is that testing and assessment, like all technologies once set in motion, proceed largely by their own momentum irrespective of the intentions of the originators (Winner 1986; Boorstin 1978). Restraints on how an existing technology is used can be very tenuous indeed (Ellul 1964, 1990). The history of testing has plenty of examples of how tests designed or mandated for one purpose have been used for another. Current proposals to make a national test or a district, school, or individual NAEP test voluntary, or even to rule out its use for high-stakes decisions, must be taken with a grain of salt. The use of the SAT as a wall-chart indicator, a criterion for athletic eligibility, a basis for awarding merit scholarships in New York, and a criterion for student-loan eligibility are all examples of a national test being used for purposes other than those originally intended, despite the objections of the CEEB and ETS. An individual NAEP will be just as susceptible to misuse as the SAT and will have greater consequences for individual students. In addition, down the road any safeguards put into place to make a district, school, or individual NAEP initially palatable will be as susceptible to change by national, state, or local policymakers as were NAEP's original procedural and technical safeguards prohibiting any individual, school, district, or state, from being singled out for description or comparison.

The disintegration of the original NAEP safeguards also illustrates how attitudes toward a national test, national standards, and a national curriculum have changed dramatically over time. How does the public feel about a federal test of a federally prescribed set of subject frameworks used to make important decisions about elementary and secondary students? Our best source of data on how a federal test and curriculum would be received is the Annual Gallup Poll of the Public's Attitude Toward the Public Schools. The Gallup survey offers an indication of how attitudes toward a national test and a national curriculum have evolved over the last two decades. Since 1970 Gallup has asked respondents:

> Would you like to see the students in the local school be given national tests, so that their educational achievement could be compared with students in other communities? (Gallup and Elam 1988, 41)

Since 1970, a majority of people have favored such a test with sentiment strongest in 1988. (The question was not asked in the 1990 survey.) However, the meaning of the term *national tests* in the Gallup question is far from clear. It might well be interpreted to mean a traditional commercially available, nationally norm-referenced achievement battery given by most local school districts with which most parents are familiar. However, the following question removes any such ambiguity and speaks directly to several current proposals for a national test:

> Should all high school students in the United States be required to pass a standard nationwide examination in order to get a high school diploma? (Gallup and Elam 1988, 40)

While one out of two respondents supported a national high school certification exam in 1958, the ratio had increased to almost three out of four the last time the question was asked in 1988. The question is interesting because of the phrases "all high-school students" and "a standard nation-wide examination," which imply a *common* examination for all students--college-preparatory, general, and vocational. Opinions about the implications of such a system, particularly on weaker students, were not explored. Further, the emphasis was on high school students. Plans to extend NAEP would presumably include elementary and middle schools students. It would be interesting to learn how the public would view using a national or federal test to make important decisions about fourth graders, for example.

The 1978 survey sheds some light on the issue of control. Respondents were asked whether tests in different subjects used for promotion should be prepared at the local, state, or national levels. Only 28 percent chose the national option, while 37 percent and 25 percent chose the local and state options respectively (Gallup 1978). In 1987, people were asked whether they favored or opposed educational achievement test results being reported on state-by-state and school-by-school bases so that comparisons could be made between schools of similar size and racial and economic make-up. (The source of the achievement tests is not made explicit, i.e., existing commercially available tests, NAEP, or some new national test.) Seventy percent were in favor of the proposal; only 14 percent were in opposition (Gallup and Clark 1987).

In 1980, respondents were asked which level of government should have the greatest influence in deciding what is taught in the local public schools. Sixty-eight percent chose local government, 15 percent state, and only 9 percent the federal government (Gallup 1980). In both 1982 and again in 1986, respondents were asked to think about the future in terms of whether the federal government should have more or less influence in determining the educational program of the local public schools. In both years, slightly over half the respondents chose the less influence option; about a quarter opted for more influence, while about one in ten opted for the same influence. In 1986, the same sort of question was asked about the relative influence of state government. Forty-five percent chose the more influence option while 32 percent chose less influence and 16 percent opted for the same as now choice (Gallup 1980).

The 1989 poll contained a series of questions on support for national goals, standards, curriculum, and testing (Elam and Gallup 1989). While past polls showed strong support for local control, the results of the 1989 poll showed that a majority of respondents supported the idea of national achievement standards and goals (70 percent); 69 percent favored a standardized national curriculum, and 77 percent favored a standardized national test. When asked who should establish the standards, the federal and state options only drew 5 percent and 3 percent respectively; the majority (61 percent) chose the professional educator option while 20 percent chose the panel of parents and other lay people (Elam and Gallup 1989). These data could be interpreted as showing a distrust of state and federal intervention in the standard-setting process.

However, in 1990 when asked about the President's and the governors' six national goals for education for the year 2000, people strongly endorsed the goals. Nonetheless, people were profoundly skeptical about the possibility that the goals can be reached within the decade (Elam 1990).

The data from the various Gallup polls would indicate support for a national test--even one for graduation decisions--and a national curriculum of some kind. However, before concluding that any national test or curriculum at any grade, for any purpose, would meet with public favor, questions would need to be more sharply drawn. For one thing, it would be wise to tease out the tension between local control on the one hand and a national test and curriculum on the other. It would also be wise to pose the question in terms of a federal test and federal curriculum. Respondents were never faced with a direct confrontation between the three concepts and most likely never considered the effects of a national or federal test or curriculum on the cherished concept of local control or even state control. It would be wise to also ask what subjects, content, and levels would be included in a national curriculum and how closely they would align with the NAEP frameworks; what the adjective *national* means (commercial, private, quasi-governmental, or governmental); what such tests would be used for (e.g., promotion or so-called gates tests), graduation, certification of the successful completion of a given level of schooling or a given type of schooling (New York State Regents), selection (college, employment), or school accountability (NAEP); what subjects, content, skills, and abilities the tests would and would not measure and at what grade(s) the tests would be administered. Further, in light of the strong sentiments in favor of local control and in favor of professional educators--not state or federal government officials--setting standards, questions about the sponsorship, control, operation, and funding of a national test and national curriculum need to be asked.

IV. Practical Problems[11]

NAEP was not designed to be a test for districts, states, and individuals, and quite properly so. Its enabling legislation precluded the possibility of reporting results for individuals, schools, districts, and, until 1990, states, and so NAEP was designed to produce national and regional results as efficiently as possible. Schools and students within schools were sampled because there was no need to assess all students. Matrix sampling and later BIB spiraling were used to sample a broad consensus of appropriate items, with the result that individual students received nonparallel booklets of often quite different difficulty. NAEP produces five "plausible values" for each student instead of individual student scores. These plausible values reflect the uncertainty of the individual results and are effective in estimating the national and regional population proficiencies that NAEP reports. The NAEP plausible values are not designed for estimating an individual student's proficiency and thus are not designed to meet professional standards for individual test scores. The NAEP plausible values are *not* appropriate for individual decision-making; they are appropriate for estimating distributions of proficiency in large populations.

One test or type of test will not fulfill all potential needs for test information, and different tests must be developed, administered, and scored for specific information needs. Different purposes imply different practical features. For instance, individual diagnostic tests are most appropriately administered in the fall, so that the teachers will have the opportunity to use the information, and accountability tests are most appropriate in the spring, after the educational system has had an opportunity to affect student learning. NAEP is generally administered in the middle of the school year, but, as a general indicator of student performance, it can be administered at any time of the year as long as the timing is consistent for comparisons over time. In timing and in many other ways, the practical design of any important test must reflect its purposes.

Given the way that NAEP has developed, a move to reporting individual results from NAEP presents special problems. The NAEP item-sampling procedures that randomly present different test booklets to different students would not be appropriate for individual student testing because the different booklets vary in difficulty and coverage.

---

[11] Much of the material in this section was drafted by Albert Beaton.

NAEP's broad area coverage is appropriate for a wide spectrum of students but is not necessarily efficient or appropriate for detailed student reporting. A national assessment can, indeed must, cover all levels of proficiency within a grade, for all types of schools, but cannot make fine distinctions. A test that is appropriate for students whether they are taking shop mathematics, business mathematics, or advanced algebra--or no mathematics at all--cannot produce detailed information about individual performance in the areas in which they are taught. A more targeted examination is necessary to allow individual students to show what they know and can do.

As mentioned above, NAEP does not measure individual students reliably. The NAEP 1990 Trial State Assessment, for example, estimated the proficiency of eighth grade students within a state on five mathematics subscales (numbers and operations, measurement, geometry, data analysis and statistics, and algebra). Using an item sampling plan, individual students were assessed for about 45 minutes and in that time presented about 42 percent of the items in the NAEP eighth grade mathematics item pool. Some students received as few as six or seven items for some subscales with the result that the reliability of individual scores on these subscales was in the .5 to .6 range, which is not adequate for individual reporting. Even if the student testing time were extended so that the students were given 105 minutes to answer all of the questions in the item pool, the reliabilities of the subscale estimates would range from about .8 to .9, still not as high as one would like for individual decision-making. The NAEP mathematics assessment would have to be enlarged if individual scores were reported and used for important decisions that affected the students.

The NAEP assessment system becomes a more appropriate estimation system as the population being estimated increases in size. NAEP reports standard errors for each estimated population characteristic. It is doubtful that the NAEP procedures would produce adequate estimates for small groups such as classrooms, schools, or small school districts. Furthermore, the NAEP assessment objectives, which are defined by a national consensus, may not echo local objectives adequately.

The NAEP assessment system is quite efficient for its national purposes, but very expensive for individual testing. NAEP's cost per student is reasonably small when compared to the alternative of testing every American student in its grade levels, but very expensive at the individual student level. The cost per assessed student of NAEP has been estimated at around $100 per student, as compared to a few dollars for a traditional standardized test. Part, but not much, of these costs would be reduced if NAEP were to attempt to assess every student. For example, the test development costs would be proportionately smaller, and the cost of sampling would be reduced. But much of the NAEP cost is simply administering and scoring the assessment, which would be increased. Using performance, portfolio, and authentic items can be easily justified when only a small sample of a population of students is assessed, but the cost and resultant turnaround delays of such items may make them unaffordable for national testing. It has been estimated that the cost of the sort of short answer open-ended item used in NAEP is about five dollars to administer and score, including the costs of training, housing, and supervising the scorers. Such costs are a heavy part of the NAEP expenses. Moreover, if a move was made toward more complex open-ended items (such as solving a complex problem in context, with individuals' performance rated live or via video tape) costs might easily escalate to $20 or more per question or task. In several European countries, for example, the cost of scoring five to seven essays from an on-demand-type examination has been estimated to be about $135 per student (Madaus 1991). Practical problems associated with non-multiple-choice items also include the amount of time needed to score results, and thus the possible usefulness of individual reports.

One of NAEP's great strengths is its national norms, which are possibly the best national educational norms ever assembled. These norms represent how students in American schools are doing on a carefully selected sample of items under a specified set of conditions. If the assessment conditions are changed, then the performance of students can be expected to change as well. If NAEP were to become a high-stakes test with consequences for schools, districts, or individual students, then the NAEP norms would not be appropriate for comparisons. Even the moderate change in NAEP that allowed individual state reports in 1990 resulted in modest but statistically significant changes in student performance so that the state data had to be equated to the national (Koffler 1991). If NAEP were to produce individual high-stakes scores to volunteering districts, there would be no technically adequate way to equate the performance of their students to those from other districts that did not request individual scores. In short, current NAEP norms would be inappropriate for interpreting individual performance on NAEP tests.

For use as a tool for school accountability, NAEP would have to be greatly modified and extended. NAEP assesses only 4th, 8th, and 12th graders and does not report on random samples of the other grades. How would the many schools that encompass the kindergarten to third grades be judged? How would the in-between grade teachers be judged? How would teachers in the 4th, 8th, and 12th grades feel about accepting responsibility for several years of growth? How would high-mobility schools be judged? Designing an appropriate test for accountability purposes would have to address these questions as well as keep track of the amount of time in which the school had the opportunity to affect the students' proficiencies. Attempting to extend NAEP for local accountability would have an enormous effect on NAEP.

Finally, by way of practical considerations, we should note that one of the fundamental features of NAEP that has contributed to its technical excellence would greatly limit its utility for purposes such as school accountability or improving student learning. Here we refer simply to the care and time that is taken between administration of NAEP instruments and the reporting of results. Since the 1970s, NAEP results have typically been reported publicly about a year or more after the instruments were administered. For example the Trial State Assessment results released in June 1991 were based on assessments actually conducted February 5 through March 2, 1990, around 15 months earlier. The delay in reporting has been used for technical analyses and for preparation of responsible interpretation of results.

Probably the most clear instance of the value of such delayed reporting occurred with regard to the reading anomaly in 1986 reading assessment results. "When they were first produced, the NAEP estimates of the reading proficiency of students in American schools in 1986 appeared anomalous in the judgment of the NAEP project staff and its technical advisors. Thus, the publication of trend results from the 1986 assessment was suspended; instead the trend results were subjected to further investigation. . . ." (Beaton and Zwick 1990, 3). Such delay seems to us to have been entirely warranted, for if results had been reported as initially derived they would have been seriously misleading.

Such delays in reporting of NAEP results in the past, it seems to us, have well served the broad purpose of NAEP of charting national and regional results over time. But note that delays of a year or more in reporting results would largely vitiate the utility of results for current school-level accountability or for monitoring district, school, and individual student progress.

V. The Problem of High-Stakes Testing

An important predicate of this paper is that if the NAEP prohibition is lifted and NAEP results are used to make high-stakes decisions about districts, schools, and individual students, NAEP's usefulness as an indicator of the nation's educational progress will be corrupted. Further, a de facto federal curriculum will evolve around the NAEP tests that teachers will teach to, and the negative effects of such practices on students, teachers, and institutions will outweigh any short term benefits. As noted above there is considerable evidence that the high-stakes testing programs of the 1970s and 80s had harmful individual, educational, and societal effects, and the evidence from other nations with high-stakes individual testing programs does not contradict the recent American experience. Perhaps the most convenient way to summarize the considerable literature on the impact of high-stakes testing on education is to list major conclusions from the literature in the form of six principles[12].

Principle 1. The more any quantitative social indicator is used for social decisionmaking, the more subject it will be to distort and corrupt the social processes it is intended to monitor.

This general principle comes from Campbell's work on social indicators (Campbell 1975) and is not limited to testing per se but extends to any social indicator used to describe, make decisions about, or influence an important social process. This principle reminds us that educational measures like NAEP are subject to a social version of Heisenberg's uncertainty principle--any measurement of the status of an educational institution, no matter how well contrived, inevitably changes its status. NAEP is a good example of this principle, and in its current form it has changed education in small but important ways for the better. However, when the measure is used for important decisions, as would be the case if NAEP were used to make high-stakes decisions about districts, schools, or individuals, then the changes would, we predict, be both large and corrupting.

Principle 2. The power of tests and examinations to affect individuals, institutions, curriculum or instruction is a perceptual phenomenon: if students, teachers, or administrators believe that the results of an examination are important, it matters

---

[12] These themes are taken from and are discussed in more detail in Madaus 1988 and in Madaus and Kellaghan, in press.

very little whether this is really true or false--the effect is produced by what individuals perceive to be the case.

Our second principle is attributable to Bloom (1969). Its importance lies in the fact that when people perceive a phenomenon to be true, their actions are guided by the importance perceived to be associated with it. The greater the stakes perceived to be linked to test results, the greater the impact on instruction and learning. A high-stakes test is one where the results are seen by students, teachers, administrators, parents, or the general public to be used to make important decisions that immediately and directly impact on individuals or institutions of concern. To date NAEP has not been perceived by important educational constituencies as directly affecting students or institutions. This is one reason why NAEP has given the nation accurate data about performance, while the high-stakes state testing programs gave misleading information about progress. However, our contention is precisely that proposed changes in NAEP contained in *AMERICA 2000* and NAGB's proposal to lift the prohibition involve high-stakes decisions and that this fact would be quickly recognized, triggering the remaining four principles. In fact, in evaluating the state-level NAEP, let alone provisions to lift the prohibition and use NAEP below the state level, one must consider whether the perception of NAEP's importance and impact on various constituencies is changing.

Principle 3: If *important* decisions are presumed to be related to test results, then teachers will teach to the test.

This principle echoes the contention of NAEP's early critics. The accommodation by teachers to a high-stakes test is often seen as having both positive and negative consequences. High-stakes tests, it is often argued, can focus instruction by giving students and teachers specific goals to attain. If the test is measuring some skill, say basic reading, preparing students for the skill measured by the mandated reading test could, proponents argue, serve as a powerful lever to improve reading skills. Unfortunately, however, the only sort of evidence generally available to bolster this proposition is that scores on high-stakes tests do tend to increase over time. People too often fail to distinguish between the skill itself and the indicator used to represent it. Repeated experience shows us that test scores can be raised in many ways without changing the levels of the skills the tests are intended to measure (for example, see Koretz, Linn, Dunbar, and Shepard 1991). One must ask what technical safeguards could be put in place to prevent a NAEP Lake Wobegon effect from happening if NAEP is extended to include district, school, and individual reporting, or if high-stakes use of NAEP results can be mandated by SEAs and LEAs?

If the test is specific to a specialized curriculum area (e.g., college preparatory physics), then the examination will eventually narrow instruction and learning, focusing only on those things measured by the tests. Indeed, this narrowing of the curriculum has been one of the enduring complaints leveled at external examinations used for the important functions of certifying the successful completion of elementary or secondary education, and admission to third level education and/or certain jobs. It is proposed to use NAEP for some of these very purposes. If this happens then we can expect that instruction will focus narrowly on the the NAEP frameworks, thus ensuring the de facto imposition of a federal curriculum.

A review of the effects of external exams on the curriculum over many years, and in several countries, indicates that faced with a choice between objectives which are explicit in the curriculum or course outline, and a different set which are implicit in the certifying examination, students and teachers generally choose to focus on the latter.[13] Why does this happen? First, there is tremendous social pressure on teachers to see to it that their students acquit themselves well on the certifying examinations. Second, the results of the examination are so important to students, teachers, and parents that their own self-interest dictates that instructional time focus on test preparation. There is no reason to believe that the very real high-stakes uses possibly associated with a district-, school-, and individual-level NAEP will be an exception to this rule.

As we have indicated, a high-stakes test can serve as a lever for the introduction of new curricular material. However, despite the ability of high-stakes examinations to help introduce new material, eventually the weight of examination precedent--the tradition of past exams--strongly influences what is taught and learned and how. Our view of the "driving" of instruction via external examination like NAEP is that in the long term the narrowing of instruction and learning that will ensue, the consequences on students and the schools, and the destruction of NAEP's usefulness as a national indicator far outweigh any initial advantages.

---

[13] See Madaus 1991, *Phi Delta Kappan*, 226-231, for detailed references.

Principle 4. In every setting where a high-stakes test operates, a tradition of past exams has developed which eventually comes to define de facto the curriculum.

Given principle 3, the question still remains: "How would teachers cope with the pressure of a high-stakes NAEP?" The answer is relatively simple. Teachers will see the kind of intellectual activity required by previous NAEP exercises and prepare the students to meet these demands. Some have argued strongly that if the skills are well chosen and the assessment truly measures them, then coaching for an exam like the new NAEP is perfectly acceptable. This argument sounds reasonable and in the short term even works. But we need to take a longer view because the argument ignores a fundamental fact of life: If the teacher's professional worth is estimated in terms of NAEP success, teachers have great incentive to produce gains. They can do this more simply by teaching test-taking strategies based on previous exercises rather than concentrating on the development of the skill or knowledge directly. The view that we can coach for the skills apart from the tradition of assessment questions is a staggeringly optimistic view of human nature which ignores the powerful pull of self-interest. It simply does not consider the long term effects of the rewards or sanctions on behavior.

Further, if NAEP is used to make high-stakes decisions such as proposed in *AMERICA 2000* and which NAGB's lifting of the prohibition could allow, the expectations and deep-seated primary agenda of students and their parents for success on NAEP will further corrupt the process. If an individual NAEP, or any national test for that matter, is perceived as important enough by parents, a commercial industry develops, outside the schools, to prepare students for it. In this country this phenomenon can be seen in the rise of commercial firms in virtually every major city selling coaching services to students for the SAT college admissions test. In Japan it is common for parents to enroll their children in special extra-study schools known as *juku* (Fallows 1987). Beginning in the 19th century a whole industry of private coaching schools called "crammers" developed in Europe to prepare students, for a fee, for high-stakes examinations (Bowler 1983). The important point about these coaching schools is not whether they are successful in preparing students for the exam; it is instead, that the public *perceives* them as helpful and is thus willing to pay for their services. Of course private coaching for an important test has always put poor children and those living in areas where such services are not readily available at a disadvantage.

Principle 5. When test results are the sole or even partial arbiter of future educational or life choices, society tends to treat test results as the major goal of schooling rather than as a useful but fallible indicator of achievement.

Of all of the effects that can be attributed to a high-stakes test like a district-, school-, or individual-level NAEP, the effect described in this principle may well be the most damaging to education. Policymakers are well aware of the high symbolic value of tests and assessments. By mandating an exam like NAEP that all schools or students must take, policymakers are seen to be making things happen. The test becomes a synecdoche for standards. The numerical scores from high-stakes tests have an objective, scientific, almost magical persuasiveness about them, that the general public and policymakers are quick to accept. Principle 5 is wonderfully illustrated in the following observation from a nineteenth century British school inspector who observed first hand the negative effects of a high-stakes testing program operating in England and Ireland which tied pupil examination results to teacher salaries:

> Whenever the outward standard of reality (examination results) has established itself at the expense of the inward, the ease with which worth (or what passes for such) can be measured is ever tending to become in itself the chief, if not sole, measure of worth. And in proportion as we tend to value the results of education for their measurableness, so we tend to undervalue and at last to ignore those results which are too intrinsically valuable to be measured (Holmes 1911, 128).

Sixty years later, Ralph Tyler echoed the same message in the 62nd NSSE Yearbook when he warned readers that society conspires to treat marks in certifying examinations as the major end of secondary schooling, rather than as a useful but not infallible indicator of student achievement.

We see the importance society places on test scores to the exclusion of other indicators in such things as the media attention to declines in SAT scores, reports that our schools score lower than those of other countries in math and science, the Education Department's wall chart that ranks states by their performance on the SAT or ACT, newspapers ranking school districts and/or schools within districts by their performance on standardized tests, the use of test results by real estate agents in selling homes; and the money spent by parents on coaching schools for the SATs. The list could go on and on. The danger of lifting the NAEP prohibition is that the exercises and the frameworks--but particularly the exercises--will come to define educational quality to the detriment of essential goals not easily measurable.

Principle 6. A high-stakes test transfers control over the curriculum to the agency that sets or controls the exam.

The agency responsible for a high-stakes test assumes a great deal of power or control over what is taught, how it is taught, what is learned, and how it is learned. This phenomenon is well understood in Europe where a system of external certification examinations, controlled by the central government or by independent examination boards, operates at the secondary level. Such a shift in power is inherent in the various proposals to extend the use of NAEP results downward from the state level. If such a proposal is enacted, a quasi-federal NAGB will have enormous control over what is taught and learned and how things are taught and learned.

In order to minimize the negative side effects associated with the impact of such tests and to maximize possible benefits, it is important that professionals understand the value and limitations that historically have been associated with their use. They also need to understand the full philosophical, political, and educational issues associated altering the gathering and reporting of NAEP data to permit inter- and intra- state, district, and school comparisons and to make comparisons between students.

VI. First Things First in Educational Policymaking

In this paper we have attempted to set out our reservations about the desirability of lifting the current prohibition on reporting NAEP results below the state level. One fundamental difficulty in lifting the prohibition is that it is unclear exactly what purposes might be served. In part II of this paper we reviewed the purposes set out to be served by the proposed interim AAT/NAEP and discussed the possible uses that might arise if the NAGB proposal to lift the prohibition against the use of NAEP tests and data reporting below the state level were enacted. We argued that the *AMERICA 2000* and NAGB proposals were fundamentally flawed in that no test, NAEP or other, can reasonably be expected to serve the span of purposes that would be associated with either or both proposals.

In part III, we briefly reviewed the reasons behind the original prohibitions against reporting NAEP results at individual, school or school district levels. Our fundamental conclusion from that review is that the issues that motivated the original prohibitions--questions about the desirability of a national curriculum, federal government influence on education, and about appropriate pedagogical practice--are alive and with us today and ought to be addressed directly, rather than roundabout via a discussion of prohibitions concerning NAEP.

In part IV, we briefly reviewed some of the practical problems implicit in developing versions of NAEP for purposes of district- and school- level accountability or for guiding instruction. Here we concluded that many of the features of NAEP that have led to its technical excellence for its current purpose of reporting on national educational progress over time mitigate against its utility for such very different purposes.

Given the assertion that lifting the prohibition against using NAEP tests and data reporting below the state level would occasion high-stakes uses of NAEP results, we reviewed in part V what we see as the fundamental problems associated with such use. We raised the following points.

1. The more any quantitative social indicator is used for social decisionmaking, the more subject it will be to distort and corrupt the social processes it is intended to monitor.

2. The power of tests and examinations to affect individuals, institutions, curriculum, or instruction is a perceptual phenomenon: if students, teachers, or administrators believe that the results of an examination are important, it matters very little whether this is really true or false--the effect is produced by what individuals perceive to be the case.

3. If important decisions are presumed to be related to test results, then teachers will teach to the test.

4. In every setting where a high-stakes test operates, a tradition of past exams has developed which eventually comes to define de facto the curriculum.

5. When test results are the sole or even partial arbiter of future educational or life choices, society tends to treat test results as the major goal of schooling rather than as a useful but fallible indicator of achievement.

6. A high-stakes test transfers control over the curriculum to the agency that sets or controls the exam.

Our fundamental conclusion from this review is quite simple. We need to get our thinking straightened out in order *to get first things first in our educational policymaking.* We ought not take for granted vital questions such as the existence or desirability of a national curriculum, the appropriate federal role in education in the United States, the specific roles that we wish testing and assessment to play in that educational policy, and most fundamentally the appropriate balance among competing aims for education in our society.

In the latter regard, we note simply that there appears to be a *significant* difference in emphasis as to the appropriate aims for education in the United States in *AMERICA 2000* released in April and another federally-sponsored report issued just two months later. While *AMERICA 2000*, and its proposal for American Achievement Tests, suggests emphasis on core academic subjects of English, mathematics, science, history and geography, the Secretary of Labor's Commission on Achieving Necessary Skills (SCANS), in its June 1991 report *What Work Requires of Schools*, suggests a quite different set of targets for schooling. It argues that "schools should teach" and "students must learn" three kinds of "foundation" skills (five "basic skills," six "thinking skills" and five "personal qualities") and five kinds of competencies (four relating to "resources," six "interpersonal," four "information," three regarding systems, and three regarding technology). Moreover the SCANS report says that these "skills" should "be assessed as part of the AMERICA 2000 agenda" (xxi). If two federally sponsored reports on education, issued just two months apart, show such sharp disagreement as to what the aims of schooling should be and what should be assessed, it seems premature, if not foolhardy, to bend NAEP to either task.

If NAEP is fundamentally changed as called for in the NAGB proposal and in *AMERICA 2000*, we are likely to undermine the utility that NAEP has shown in the past in informing educational policymaking. More broadly, if we abandon NAEP as an instrument for *informing* educational policy, by turning it into an instrument for *making* educational policy, we are likely, as the history of high-stakes testing clearly shows, to undermine teaching and learning in our nation's schools.

The fate now threatening NAEP, it seems to us, is highly analogous to that which has befallen commercial standardized achievement test batteries, such as the Metropolitan and Comprehensive Tests of Basic Skills. For decades these tests were used quite successfully for the modest jobs of informing school people and parents about how children were performing in broad areas such as reading comprehension and math computations. But with the education reform and accountability movements of the 1970s and 80s, these instruments were widely applied to tasks which they could not (and we would argue, should not) bear, such as controlling grade promotion and school personnel evaluation. As a result, these tests, quite apart from the intentions of their creators, became clear targets for instruction, their validity undermined, and their credibility now widely suspect.

This is, we fear, exactly the sort of fate that may befall NAEP, if it is redirected away from its charter for the last 20 years, of charting educational progress nationally, regionally, and now at the state level, toward district, school and individual uses of results. If we wish, as national educational policy, for tests to play new roles in our schools, we should first clarify exactly what those roles should be, and then set about developing and validating instruments for those roles.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1985. *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

American Educational Research Association. 1991. *Researchers Say, Slow Down on National Test! - States Provide Window on Likely Failure of National Testing to Achieve Accountability and Reform Goals*. Washington, DC.

American Educational Research Association. 1991. *AERA Speaks Out Against National Test*. Washington, DC: OIA info memo.

Archer, P. 1979. "A Comparison of Teacher Judgments of Pupils and the Results of Standardized Tests." University College Cork, Ireland: unpublished masters thesis.

Bangert-Drowns, R. L., Kulik, C., Kulik, J. A., and Morgan, M. 1991. "The Instructional Effect of Feedback in Test-like Events." *Review of Educational Research,* 61(2): 213-238.

Beaton, A. E., and Zwick, R. 1990. *The Effects of Changes in the National Assessment: Disentangling the NAEP 1985-86 Reading Anomaly*. Princeton, NJ: Educational Testing Service.

Beymer, L. May 1966. "The Pros and Cons of the National Assessment Project." *Clearing House,* 40, 5: 540-543.

Bloom, B. S. 1969. "Some Theoretical Issues Relating to Educational Evaluation: New Rules, New Means." *The Sixty-Eighth Yearbook of the National Society for the Study of Education. Pt. II.* Ed. R.W. Taylor. Chicago: University of Chicago Press.

Boorstin, D. J. 1978. *The Republic of Technology*. New York: Harper & Row Publishers, Inc.

Bowler, R. F. 1983. *Payment by Results: A Study in Achievement Accountability*. Boston College.

Brody, H. 1991. "Great Expectations: Why Technological Predictions Go Awry." *Technology Review,* 94(5): 8-45.

Campbell, D. T. 1975. "Assessing the Impact of Planned Social Change," in *Social and Public Policies: The Dartmouth/OECD Conference*. Hanover, NH: Public Affairs Center, Dartmouth College.

Cannell, J. J. 1989. *The 'Lake Wobegon' Report: How Public Educators Cheat on Standardized Achievement Tests*. Albuquerque, NM: Friends for Education.

Catterall, J. S. 1989. "Standards and School Dropouts: A National Study of Tests Required for High School Graduation." *American Journal of Education, 98*(1): 1-34.

Cooley, W. W., and Bernauer, J. A. 1991. "School Comparisons in Statewide Testing Programs," in *Advances in Program Evaluation*. Ed. R.E. Stake. Greenwich, CT: JAI Press.

Darling-Hammond, L. 1991. "The Implications of Testing Policy for Educational Quality and Equality," in *American Educational Research Association Invitational Conference on Accountability as a State Reform Instrument: Impact on Teaching, Learning, Minority Issues and Incentives for Improvement*. Washington, D.C..

Educate America Inc. 1991. *An Idea Whose Time has Come; A National Achievement Test for High School Seniors!* Morristown, NJ.

Elam, S. M., and Gallup, A. M. 1989. "21st Annual Gallup Poll of the Public's Attitudes Toward the Public Schools." *Phi Delta Kappan*, 71(1): 41-57.

Elam, S. M. 1990. "22nd Annual Gallup Poll of the Public's Attitudes Toward the Public Schools." *Phi Delta Kappan*, 72(1): 41-53.

Ellul, J. 1964. *The Technological Society*. New York: Vintage Books.

Ellul, J. 1990. *The Technological Bluff*. Grand Rapids, MI: Williams B. Eerdmans Company.

Equal Employment Opportunity Commission, U.S. Department of Labor, and Office of Personnel Management. 1985. *Uniform Guidelines on Employment Selection Procedures*. Washington, DC: USGPO, 29 CFR Part 1607.

Fallows, J. 1987. "Gradgrind's Heirs." *Atlantic*: 16-24.

Fetler, M. E. 1991. "Pitfalls of Using SAT Results to Compare Schools." *American Educational Research Journal*, 28(2): 481-91.

Finn, C. E. J. 1991. *We Must Take Charge: Our Schools and Our Future*. New York: Free Press.

Gallup, G. 1978. "The 10th Annual Gallup Poll of the Public's Attitudes Toward the Public Schools." *Phi Delta Kappan*, 60(1): 34-46.

Gallup, G. 1980. "The 12th Annual Gallup Poll of the Public's Attitudes Toward the Public Schools." *Phi Delta Kappan*, 62: 3-46.

Gallup, A. M., and Clark, D. L. 1987. "The 19th Annual Gallup Poll of the Public's Attitudes Towards the Public Schools." *Phi Delta Kappan*, 69(1): 17-30.

Gallup, A. M., and Elam, S. M. 1988. "The 20th Annual Gallup Poll of the Public's Attitudes toward the Public Schools." *Phi Delta Kappan*, 70(1): 33-46.

Greenbaum, W., Garet, M. S., and Solomon, E. R. 1977. *Measuring Educational Progress: A Study of the National Assessment.* New York: McGraw Hill.

Hall, J. L., and Kleine, P. F. April 17, 1989. *Educator Perceptions of National Achievement Test Use and Abuse: A National Survey.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, MA.

Hand, H. C. 1965. "Integrity and Instructional Innovation." *The Educational Forum*, 30(1), 7-16.

Hand, H. C. 1965. "National Assessment Viewed as the Camel's Nose." *Phi Delta Kappan*, XLVII(1): 8-13.

Hand, H. C. 1966. "Recipe for Control by the Few." *The Educational Forum*, 30(3): 263-272.

Haney, W. M. 1984. "Testing Reasoning and Reasoning about Testing." *Review of Educational Research*, 54(4): 597-654.

Haney, W. October 1985. "Making Testing More Educational." *Educational Leadership*, 43(2): 4-13.

Haney, W., and Madaus, G. F. 1989. "Searching for Alternatives to Standardized Tests: Why, Whats, and Whethers." *Phi Delta Kappan*, 70(9): 683-687.

Hastings, J., Runkel, P., Damrin, D., Kane, R., and Larsen, G. 1960. *The Use of Test Results.* Urbana, IL: Bureau of Educational Research, University of Illinois.

Hazlett, J. A. 1974. *A History of the National Assessment of Educational Progress, 1963-1973; A Look at Some Conflicting Ideas and Issues in Contemporary American Education.* Lawrence, KS: University of Kansas.

Holmes, E. G. A. 1911. *What Is and What Might Be: A Study of Education in General and Elementary in Particular.* London: Constable.

Hopkins, K. D., George, C. A., and Williams, D. 1985. "The Concurrent Validity of Standardized Achievement Tests by Content Area Using Teachers' Ratings as Criteria." *Journal of Educational Measurement*, 22(3): 177-82.

Jaeger, R. M. June 5, 1991. "Legislative Perspectives on Statewide Testing: Goals, Hopes, and Desires," in *American Education Research Association Invitational Conference on Accountability as a State Reform Instrument: Impact on Teaching, Learning, Minority Issues and Incentives for Improvement.* Washington, D.C.: 14.

Kellaghan, T., Madaus, G. F., and Airasian, P. W. 1982. *The Effects of Standardized Testing.* Boston, MA: Kluwer-Nijhoff Publishing.

Kellaghan, T. November 18, 1991. "Issues in Educational Assessment in Developing Countries." Paper presented at E.T.S. Seminar for World Bank. Princeton, NJ.

Koffler. S.L. 1991. *The Technical Report of NAEP's 1990 Trial State Assessment Program* (NAEP report 21-ST-01). Princeton, NJ: ETS.

Koretz, D., Linn, R., L, Dunbar, S. B., and Shepard, L. A. 1991. "The Effects of High-Stakes Testing on Achievement: Preliminary Findings about Generalization across Tests," in *Annual Meeting of the American Research Association and the National Council on Measurement in Education*. Ed. D. M. Koretz. Chicago: The Rand Corporation.

Kreitzer, A. E., Haney, W., and Madaus, G. F. 1989. "Competency Testing and Dropouts," in *Dropouts from School: Issues, Dilemmas, and Solutions, Part II*. (Ed. L. Weis, E. Farr, and H. G. Petrie. Albany, NY: State University of New York Press.

Madaus, G. F., and Greanery, V. 1985. "The Irish Experience in Competency Testing: Implications for American Education." *American Journal of Education*, 93(2): 268-294.

Madaus, G. F. 1988. "The Influence of Testing on the Curriculum," in Critical Issues in Curriculum, *Eighty-seventh Yearbook of the National Society for Study of Education*. Ed. L. Tanner. Chicago: University of Chicago Press.

Madaus, G. F., and Kellaghan, T. 1990. *Curriculum Evaluation and Assessment*. New York: Macmillan.

Madaus, G. F., Kellaghan, T. 1991. *Student Examination Systems in the European Community: Lessons for the United States*. Contractor Report submitted to the Office of Technology Assessment, United States Congress.

Madaus, G. M. 1991. "The Effects of Important Tests on Students: Implications for a National Examination or System of Examinations." *Phi Delta Kappan*, 73(3):226-231.

Madaus, G. F., and Kellaghan, T. (in press). "Curriculum, Evaluation and Assessment," in *Handbook of Research on Curriculum*. Ed. P. W. Jackson. Washington, DC: American Educational Research Association.

Malone, T. W. 1980. *What Makes Things Fun to Learn? A Study of Intrinsically Motivating Computer Games*. Palo Alto, CA: Xerox Palo Alto Research Center.

Metropolitan District Commission. 1988. *America's Shame, America's Hope*. Chapel Hill, NC: Charles Stewart Mott Foundation.

Miller, J. A. April 24, 1991. "Bush Strategy Launches 'Crusade' for Education." *Education Week*: 1,26.

National Assessment Governing Board. November 29, 1989. "*The Future of NAEP as the Nation's Report Card*. Washington, DC. (Available from NAEP).

National Assessment Governing Board. December 9, 1989. *Positions on the Future of the National Assessment.* Washington, DC. (Available from NAEP)

National Assessment Governing Board. November 17 1990. *Positions of the National Assessment Governing Board on the 1994-1996 NAEP.* Atlanta, Georgia. (Available from NAEP)

National Education Goals Panel. 1991. *Measuring Progress toward the National Education Goals: Potential Indicators and Measurement Strategies.*

Natriello, G., and Dornbush, S. M. 1984. *Teacher Evaluative Standards and Student Effort.* New York: Longman.

O'Neil, J. 1991. "Drive for National Standards Picking Up Steam." *Educational Leadership,* 48(5): 4-9.

Paris, S. G., Lawton, T. A., Turner, J. C., and Roth, J. L. 1991. "A Developmental Perspective on Standardized Achievement Testing. *Educational Researcher,* 20(5): 12-20.

Pedulla, J. J., Airasian, P. W., and Madaus, G. F. 1980. "Do Teacher Ratings and Standardized Test Results of Students Yield the Same Information?" *American Educational Research Journal,* 17: 303-7.

Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D., and Williams, P. L. May 1985. "Measurement-Driven Instruction: It's On the Road." *Phi Delta Kappan,* 66(9): 628-635.

Powell, B., and Steelman, L.C. 1984. "Variations in State SAT Performance: Meaningful or Misleading?" *Harvard Educational Review, 54:* 389-412.

Prost, A. 1968. *Histoire de l'enseignement en France* 1800-1967. Paris: Colin.

Radwin, E. 1981. *A Case Study of New York City: Citywide Reading Testing Program.* Cambridge, MA: The Huron Institute.

Rafferty, M. 1985. "Examinations in Literature: Perceptions from Nontechnical Writers of England and Ireland from 1850 to 1984." Boston College: unpublished doctoral dissertation.

Raudenbush, S. 1984. "Magnitude of Teacher Expectancy Effects on Pupil IQ as a Function of the Credibility of Expectancy Induction." *Journal of Educational Psychology,* 76(1): 85-91.

Resnick, L. 1991. *An Examination System for the Nation.* University of Pittsburgh: 1-9.

Rotberg, I. C. 1990. "I Never Promised You First Place." *Phi Delta Kappan,* 72(4): 296-303.

Rotberg, I. C. June 1991. "How Did All Those Dumb Kids Make All Those Smart Bombs?" *Phi Delta Kappan*, 72(10): 778-781.

Salmon-Cox, L. 1981. "Teachers and Standardized Achievement Tests: "What's Really Happening?" *Phi Delta Kappan*, 62: 631-634.

Saylor, G. 1970. "National Assessment: Pro and Con. *Record*, 71(4): 588-597.

Shanker, A. 1990. "Perestroika for the Classroom." *New Perspectives Quarterly*, 7(4): 20-22.

Shanker, A. February 24, 1991. "An American Revolution in Education: Developing a Common Curriculum." *New York Times:* E-7.

Shepard, L. A. June 5, 1991. "Will National Tests Improve Student Learning?" in *American Education Research Association Invitational Conference on Accountability as a State Reform Instrument: Impact on Teaching, Learning, Minority Issues and Incentives for Improvement*. Washington, DC.

Smith, M. S., O'Day, J., and Cohen, D. K. Winter 1990. "National Curriculum, American Style: Can It Be Done? What Might It Look Like?" *American Educator*, 14(4): 10-17, 40-47.

Stake, R. E. June 5, 1991. "The Teacher, Standardized Testing, and Prospects of Revolution," in *American Education Research Association Invitational Conference on Accountability as a State Reform Instrument: Impact on Teaching, Learning, Minority Issues and Incentives for Improvement*. Washington, DC.

Thorndike, E. L. 1913. *Educational Psychology*. New York: Teachers College.

Tyler, R. W. 1965. "Assessing the Progress of Education." *Phi Delta Kappan*, 47(1): 13-16.

U.S. Department of Education. 1991. *AMERICA 2000 Sourcebook*. Washington DC: U.S. Department of Education.

U.S. Department of Education. 1991. *AMERICA 2000: An Education Strategy* (ED/OS91-13). Washington, DC: U.S. Department of Education.

Winner, L. 1986. *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. Chicago: University of Chicago Press.

Reasonable Inferences for the Trial State NAEP Given the Current Design: Inferences That Can and Cannot Be Made

Edward H. Haertel
Stanford University

October 1991

Acknowledgement

I wish to express my appreciation for the support and willing assistance of the staffs of the National Academy of Education and the American Institutes for Research. Gene Johnson and Robert J. Mislevy, both of the Educational Testing Service, provided helpful insights and technical clarifications. Finally, I am indebted to the reviewers of this paper, including George Bohrnstedt, Robert Glaser, Robert Linn, Gordon Ambach, and especially Linda Darling-Hammond, whose many marginal comments provoked much thought and led to numerous improvements. Any errors or lapses of logic, of course, remain my own responsibility.

Abstract

The 1990 and 1992 Trial State Assessments (TSAs), conducted as part of the National Assessment of Educational Progress (NAEP), represent a major shift in the structure and purposes of the national assessment. For the first time, state-level achievement data will be available through NAEP, permitting comparisons of states to one another and to the nation as a whole. This paper has been commissioned as part of the evaluation of the TSAs being conducted by the National Academy of Education and the American Institutes for Research. It reviews the structure of the data that will become available from the TSAs, offers a technical critique of their strengths and limitations, and considers a range of inferences that may be attempted using these data. It is argued that the most useful and defensible inferences will be primarily descriptive, documenting absolute levels of achievement, relative and absolute performance of major student subgroups, trends over time, and, from the student, teacher, and principal background questionnaires, documenting instructional resources and practices for different student subgroups and over time. The TSAs may also suggest hypotheses for further study and may help to keep educational issues in the public eye. The TSA data are not likely to support causal inferences about the effectiveness of specific state-level policies or practices, although trend lines over several years could certainly inform the evaluation of major, state-level educational reforms. In addition, focused comparisons of particular subgroups across states with similar demographic characteristics might directly inform the efficacy of educational policies and practices.

Introduction

The expansion of the National Assessment of Educational Progress (NAEP) to enable state-level reporting in eighth grade mathematics (1990) and later in fourth and eighth grade mathematics as well as fourth grade reading (1992) marks a signal departure from the original conception of the assessment and even from its original purposes. After nearly two decades during which NAEP was conceived almost exclusively as an unobtrusive, nonreactive instrument for monitoring national achievement trends, carefully designed to preclude reporting below the level of broad geographic regions, it has the potential to become, in a matter of months, the most authoritative source of information on both absolute and relative levels of educational achievement across the states.

The information available from the 1990 and 1992 Trial State Assessments (TSAs) will be limited to just two grade levels and two subject areas. Moreover, in 1990 only 37 states, 2 territories, and the District of Columbia chose to participate, and in 1992 41 states, 2 territories, and the District of Columbia plan to participate. Nonetheless, there is every reason to expect that results of the TSAs will be widely publicized, used to criticize or defend a broad array of educational policies, and closely scrutinized by educational researchers and methodologists. The TSAs could potentially influence real estate values, corporate decisions about the siting of new factories, education budgets, curriculum frameworks, educational policies, and political careers. It is not at all clear what the range of *intended* uses and consequences of the TSA information may be, let alone unintended consequences.

The TSAs will yield two broad categories of information. Perhaps most important, they are serving as trials of new assessment methods and are providing invaluable information about the methods and technical characteristics of such assessments themselves. The second kind of information, also very important, will be conveyed by the responses of students, teachers, and principals to the background questions and cognitive exercises they are presented. This paper focuses on the second kind of information. On the basis of the limited information and incomplete reporting plans now available, the technical quality of the information the TSAs may be expected to yield is addressed and a range of both warranted and unwarranted inferences that may be made from this information is considered.

The paper begins with a brief discussion of the policy context of the TSAs. At the conclusion of this section, I summarize some positive uses of the TSA data, before turning to a more detailed critique in the body of the paper. The next major section reviews the general structure and the contents of the database the TSAs will yield, then considers issues in sampling, in data collection, in the NAEP exercise pool itself, and in the scaling and reporting of the TSA findings, including some discussion of performance standards that may be established. The third major section describes a series of specific inferences that may be attempted and considers the degree of confidence that may be placed in each type of decision. The paper concludes with a brief summary and bulleted recommendations.

Policy Context

Over the past 5 to 10 years, a strong demand has developed for better state-level achievement data. Increases in state-level funding of education began well over a decade ago with the school finance reform initiatives of the 1970s, and in the 1980s, Reagan's "new federalism" decentralized responsibility for administration of much federal educational funding, further increasing the states' roles in educational management. The dramatic educational reforms enacted subsequently by many states brought heightened expectations for improved achievement test scores together with heightened pressure for accountability, for tangible evidence that more dollars were bringing better results. (Major state-level reform initiatives continue up to the present, e.g., ongoing school restructuring in Kentucky and in Oregon.)

Also beginning in the early 1980s, the Secretary of Education issued annual "wall charts" ranking the states according to various educational indicators, with achievement shown only by average Scholastic Aptitude Test (SAT) or American College Test (ACT) scores, whichever test was taken by more examinees in a given state. The inadequacy of these scores, from self-selected samples taking tests designed for a very different purpose, drew attention to the absence of any better data on educational outcomes at the state level. The National Governors Association (NGA) and the Council of Chief State School Officers (CCSSO) both issued calls for more accurate and comparable state-level achievement data. Thus, it came as no surprise when in 1987 the National Academy of Education's study group report on NAEP, *The Nation's Report Card,* called for an expansion of the national assessment to enable state-level reporting (Alexander and James 1987). In 1988, with the passage of PL 100-297 (the Hawkins-Stafford Act), Congress authorized the expansion of NAEP to permit collection and reporting of data at the state level in 1990 and 1992, and the TSAs were initiated.

The policy context for reports and interpretations of NAEP data, including the TSAs, is changing rapidly. The education goals announced by the President and taken up by the governors at Charlottesville presuppose that adequate measures of educational achievement exist or will be created. To the extent that the states accept responsibility for implementing and attaining national goals relating to measured achievement, NAEP is likely to assume unprecedented importance.

In a related development, the National Assessment Governing Board (NAGB) has announced plans to establish three performance standards at each of the fourth, eighth, and twelfth grade levels in each NAEP subject area, representing basic, proficient, and advanced levels of attainment. Unlike the scale-point descriptions developed by the Educational Testing Service (ETS, the present NAEP contractor), the NAGB standards are to be explicitly normative, with the "proficient" standard representing a level of solid accomplishment that ought to be expected of students ready to progress to the next grade level. The process of establishing standards for the eighth grade mathematics assessment is already underway.

On paper, the states participating in the 1990 TSA will have considerable latitude in deciding how their own results will be released and whether they will be interpreted with respect to the NAGB standards. However, it is likely that every participating state will face strong pressure to permit reporting of the proportions of its eighth graders attaining the "basic," "proficient," and "advanced" achievement levels in mathematics. This form of reporting will invite inferences about the "absolute" levels of educational attainment in each state and in the nation, as well as the "absolute" quality of the educational system. Such inferences will take their place alongside "relative" comparisons among states, among subpopulations, or across different points in time, and will also have important consequences. Obviously the validity of such "absolute" inferences depends critically on the adequacy of the standard-setting process and also on the accuracy with which the NAEP exercise pool can discriminate achievement levels in the regions of the cutting scores representing each standard, among many other substantive and technical considerations.

Finally, the second "International Assessment of Educational Progress" (IAEP-91) will raise the stakes in NAEP reporting and invite a still broader range of inferences from NAEP data by enabling links from NAEP mathematics and science scales to tests used for broad cross-national comparisons. In principle, IAEP-91 together with the TSAs could make possible achievement comparisons between the eighth grade students in a given state and students of a comparable age in various nations around the world. Obviously, a full consideration of the validity of such comparisons would entail close examination of IAEP-91 and the national component of NAEP as well as the TSAs, and would go well beyond the scope of this paper.

Justifiable Uses and Interpretations of the TSA Data

In one sense, findings from the TSA are likely to offer few surprises. Overall rankings of the participating states will be highly predictable from readily available state-level data including parent's educational level, household income, average class size, per-pupil expenditures, and other demographic and educational variables.[1] Despite the predictability of overall rankings, the TSA data may yield important information about absolute levels of achievement, about the relative and absolute performance of important subgroups of students, about instructional resources and practices, and about trends over time. For reasons set forth in the detailed discussion to follow, I believe that the most important and defensible inferences from the TSA data will be those that minimize theorizing about causes and effects, focusing instead on the accurate description of schooling practices and learning outcomes. In participating states where these groups are well represented, the TSA data can document with unprecedented precision what fractions of white, black, and Hispanic boys and girls in the public schools of our cities, suburbs, and rural areas are capable of solving particular types of exercises. TSA background questions addressed to students, teachers, and principals can begin to document characteristics of the instruction offered these different student groups and can offer a rich source of hypotheses about the effects of policies and practices, testable in more focused studies.

---

[1] I do not mean to suggest that these particular variables, alone or in combination, are the most highly predictive of achievement test scores. They are merely illustrative.

As with any large-scale assessment, TSA outcomes, trend lines, and subpopulation comparisons will always have to be interpreted cautiously. Improvements over time in a state's achievement levels may reflect broadly improved educational outcomes or may merely reflect a tuning of the curriculum to the test, shifting topics so that those covered on, say, the fourth grade test are taught before it is administered and other topics not tested are deferred until later. As NAEP test scores become more important, teachers may incorporate worksheets into their ongoing instruction with exercises that resemble those in the assessment, leading to better scores but not to any genuine educational improvement. Over time, students from higher and lower socioeconomic levels may migrate differentially between public and private schools, influencing public-school score distributions in ways that have nothing to do with levels of learning outcomes. Eligibility criteria for inclusion in the NAEP testing may change or be interpreted differently, affecting the population tested in ways that are difficult to discern. A reduction over time in the achievement gap between two subgroups in a given state may be the result of some mix of educational policies, but may also reflect changing demographics (e.g., the effects of large-scale, between-state migrations or changing immigration patterns).

Despite these and many similar cautions, state-level assessments can provide increasingly valuable achievement trend lines for states and for subpopulations within states, especially if they are continued beyond the 1990 and 1992 TSAs and expanded to more content areas and to all three age/grade levels. Several years of NAEP data could provide an extremely valuable baseline against which to evaluate the global effects of major educational restructuring efforts at the state level. Comparisons between carefully matched student subgroups in selected states with similar demographics could offer compelling evidence that better schooling outcomes were attainable. Cautiously interpreted, trends over time in the achievement disparities between, say, blacks and whites or Hispanics and whites could indicate progress or lack of progress in resolving long-standing inequities in schooling outcomes.

Another important function of the TSA may be to help drive home the message that concerns expressed about American education are addressed to local schools, including those near at hand. The NAEP proficiency scales and released exercises can help to make concrete the levels of knowledge and skill demonstrated by children in each participating state, lending urgency to calls for educational reform and keeping educational issues in the public eye.

The National Assessment of Educational Progress represents the state of the art in large-scale achievement testing. Over the years, NAEP has been a major source of technical progress and innovation and has served as a model for large-scale assessment programs at the state level as well as recent international assessments. The TSAs continue a tradition of excellence. Despite some compromises and limitations, they offer a rich resource for tracking educational outcomes and documenting instructional practices. The purpose of this paper is not to denigrate the TSAs, but merely to draw attention to their strengths while discouraging unwarranted over-interpretations.

Structure and Technical Characteristics of the TSA Data

This section describes the overall structure of the data the 1990 and 1992 TSAs may be expected to yield, and discusses their strengths and weaknesses. Of necessity, the discussion is at points somewhat speculative. The 1990 TSA data are only now being analyzed and the 1992 TSA is only now being designed; the NAGB standard-setting effort is not yet complete; and at most preliminary results are available from the various ongoing evaluations and special studies that bear on the TSAs.[2] Issues in this section are reviewed under six headings: the TSA data structure; the contents of the TSA database; sampling; data collection; exercise pool; and scaling, analysis, and reporting.

---

[2] Congress has mandated an evaluation of the TSAs to determine "the feasibility and validity of [state] assessments and the fairness and accuracy of the data they produce." This evaluation is being conducted by the National Academy of Education in collaboration with the American Institutes for Research. A separate NAEP Technical Review Panel (TRP) has been constituted under the direction of the Center for Research and Evaluation on Standards and Student Testing (CRESST) and the RAND Corporation. The charge of the TRP encompasses more than the TSAs, but several of its studies will directly inform the quality and usefulness of the TSA data. Yet another ongoing evaluation addresses the NAGB standard-setting effort, which also bears on the validity of inferences from the TSA data.

TSA Data Structure

Even though the major purpose of the national component of NAEP has been and remains the tracking of achievement trends over time, the major portion of each biannual assessment is cross-sectional. Each assessment includes nationally representative "bridge" samples, which are assessed following as closely as possible the procedures of some earlier year, right down to the color of ink in the exercise booklets. This permits the tracking of achievement trends over time. Separate, larger, nationally representative "cross-sectional" samples of fourth, eighth, and twelfth graders are also drawn for a more innovative assessment. The cross-sectional samples are large enough to permit accurate estimation of performance distributions for subgroups defined by gender, race-ethnicity, size and type of community, socioeconomic level, and other variables. For the TSAs, additional samples of about 2,000 students are drawn in each participating state. The assessment for these state samples closely parallels the national cross-sectional assessment.

The scope of the TSA database and data collection is enormous. In 1990, the TSA involved about 3,900 schools, each with its own local administrator (LA), and a total of about 100,000 students. It generated about 1,835,000 open-ended responses, which will require hand scoring. In 1992, basing projections on participation by 45 states and taking account of the expansion to include fourth grade reading and mathematics, the TSA will involve 10,500 schools and the same number of LAs, 360,000 students, and 5,750,000 open-ended responses (Mullis 1989; Mullis et al. 1990). Separately, each state sample of about 2,000 students in perhaps 100 schools is too small to support extensive within-state analyses of subpopulations. Taken together, though, the TSA samples far exceed the size of national NAEP.

Data collected on a subset of the states at a single point in time may be useful for descriptive purposes, and can obviously be used to compare those states to one another. Careful linkage to the national NAEP also enables comparisons of each participating state to the nation as a whole. In order to interpret these comparisons, however, additional structure must be created or imposed on the TSA database. For example, "co-statistics" might be introduced. A *co-statistic* is some additional variable observed for each state, which might be used to classify the states in some way or to provide some context for interpreting achievement patterns across states. Additional structure might also be introduced by distinguishing student subpopulations within states.

One way of structuring the TSA data will be to divide the states into groups according to some characteristic. This leads to what may be termed the "natural experiment" design. Just to illustrate the idea, states might be classified as Coastal versus Inland, and these groups might be compared to determine whether proximity to a large body of water (the co-statistic) facilitated achievement. (Obviously, my illustration is intended to encourage skepticism. How about comparing states with Republican versus Democratic governors, or those with versus without a common state-wide minimum competency test?)

Another approach, also using co-statistics, might be termed the "search for effective schools" design. Measures reflecting socioeconomic status, parent educational level, and other student demographic variables would be used as predictors of achievement at the state level, perhaps via multiple linear regression.[3] Those states in which achievement markedly exceeded the level predicted would then be studied closely to try and discover the key to their unusual success. Clearly, with only 37 data points (37 states) to begin with, any analyses relying on complex multivariate methods will be highly problematical. It is extremely unlikely that the enormous diversity of the states can be captured well enough in few enough co-statistics to permit meaningful outlier detection. Moreover, even if one or more states with anomalous achievement levels could be isolated reliably, the task of explaining such anomalies would be at least as formidable. Explanations might go far beyond schooling policies and practices, involving, for example, prenatal and early childhood health care services, the provision of family supports in connection with welfare and employment, or immigration patterns that challenge the schools of some states more than others to enculturate as well as educate.

Another way of introducing additional structure into the TSA database is to subdivide each state sample. For example, means could be calculated for white non-Hispanic, black non-Hispanic, and Hispanic groups within each participating state to examine the magnitude of racial-ethnic achievement disparities and how they varied from one state to another. Note, however, that for many possible ways of classification, the data matrix formed in this manner will be incomplete. The TSA samples were designed to yield accurate state-level achievement levels, and so the representation in some states of blacks or Hispanics, for example, will likely be insufficient to obtain an accurate mean achievement level. Subsample achievement estimates would be more accurate if states were pooled within geographic regions, but this would result in even fewer data points, and connections to policies implemented at the state level would be even murkier.

---

[3] It is important that regression coefficients from such aggregate-level analyses not be interpreted as estimating individual-level effects. This is the problem of so-called "ecological correlation." The fact that individual-level and aggregate-level effects differ, and that aggregate data can generally not be used to inform individual-level effects, does not in itself invalidate the aggregate-level analyses outlined here.

These and other ways of structuring the data might be used in concert, giving rise to an array of increasingly complex designs. For example, the magnitude of racial-ethnic gaps might be estimated more accurately by introducing socioeconomic co-statistics as control variables, and then related to the presence versus absence of some state-level compensatory education policy. It is well to bear in mind that more complex designs are likely to rely on stronger assumptions, especially an assumption that models are fully specified, i.e., that nothing important has been left out of the equation. Given the complex policy environment in which particular educational reforms are undertaken, such efforts are likely to be seriously misleading.

It would be possible to apply the cross-sectional approaches described thus far to the 1990 grade 8 mathematics data, the 1992 grade 8 mathematics data, 1992 grade 4 mathematics data, or the 1992 grade 4 reading data. In addition, for grade 8 mathematics, stronger designs are possible. This is the only subject-grade combination for which the TSAs authorized to date will yield two waves of data. The 1990 data were based on 7 blocks of cognitive items. In 1992, present plans call for 13 blocks of cognitive items, including 3 blocks from among the 7 used in 1990. The 1992 exercises will be based on the same content-by-process objectives grid as was used in 1990 and should be "close cousins" to the 1990 exercises, although "higher-order thinking" items will be more heavily represented. Administration procedures in 1992 should be nearly identical to those in 1990. Thus, the 1992 and 1990 proficiency scales should represent much the same dimension of achievement and should be equatable via the three common exercise blocks (Gene Johnson, ETS—personal communication 20 February, 1991). This form of linkage relies on stronger psychometric assumptions and is less accurate than the "identical administration" (bridge sample) method used to maintain national trends. Nonetheless, for those states participating in both the 1990 and 1992 TSAs, it will provide two waves of data in the same metric, so that (state-level) change over time can be examined. Note that it will not be possible to examine change over time at the level of students, classrooms, or schools. It will be possible to examine change for selected subpopulations, e.g., eighth grade boys in California.

Designs using two waves of data are somewhat parallel to single-wave designs, but with each state "serving as its own control," stronger inferences become possible. For example, states might be classified into those with "old math" versus "new math" curriculum frameworks, and 1990-to-1992 change scores could be compared for these two groups. Alternatively, states could be rated according to the degree of overlap between the TSA exercise battery and their state curriculum framework in mathematics, and the (continuous) rating variable could be correlated with their change scores. Two waves of data are better than one, but the analysis of change scores brings other problems to the fore. Eighth grade cohorts in 1990 versus 1992 might differ not only due to educational reforms during the intervening two years, but also due to between-state migrations, like the massive "rust belt to sun belt" migration during the 1980s. Demographic changes and differential educational policies affecting lower grade levels in earlier years will also influence the eighth grade change scores.

Contents of the TSA Data Base

Each student sampled for the TSA is tested for about an hour. The testing session includes a 5-minute general background questionnaire, a 5-minute subject-specific questionnaire, and three separately timed 15-minute blocks of cognitive exercises. In addition, each participating student's teacher is identified and asked to complete a teacher questionnaire, which is linked to the student data. Finally, a school characteristics and policies questionnaire (SCPQ) is completed by the principal or someone the principal designates. Other data collection instruments include an excluded student questionnaire (one for each student sampled but deliberately not tested due to language background or some handicapping condition) and various questionnaires, for local administrators, monitors, and others, relating to the conduct of the assessment itself.

The general questions address student race and ethnicity, language spoken in home, parent educational levels, homework, attendance, school climate, parent employment, whether parents live in the home, and other variables. Subject-specific questions concern instructional materials and activities, course taking, out-of-school experience with the content area, attitudes toward the subject, and other variables. Teacher questions ask about their own background and characteristics (sex, race-ethnicity, certification, years of experience, etc.) as well as their instructional practices and include some questions that closely parallel those asked of students. The SCPQ includes questions about the principal, school enrollment and attendance, and a range of school-level policies and potential school problems (Mullis et al. 1990).

The heart of each assessment, of course, is its cognitive exercises. These are grouped into blocks to be administered during some standard time duration, e.g., 15 minutes. The blocks are combined into booklets in such a way that any two blocks appear together in at least one booklet. This permits estimation of the covariance between any two exercises, regardless of which blocks they occur in. For the 1990 TSA in eighth grade mathematics, the pool consisted of seven blocks of exercises, and each student was exposed to three blocks. Seven different booklets were prepared, in such a way that each block was paired once with each other block, and each block occurred once in each of the first, second, and third positions within some booklet. Booklets were then interleaved in a predetermined cyclical order, in such a way as to assure that nearly identical proportions of students within each testing site respond to each booklet. The name given this overall design is balanced incomplete block spiraling, or BIB-spiraling. (It is sometimes referred to as *focused-BIB spiraling* to indicate that the BIB design is implemented within, not across, content areas.) The national NAEP in eighth grade mathematics included 10 blocks of cognitive items, of which only 7 were used in the TSA. The three omitted blocks included most of the "estimation" and "higher-order thinking" exercises.[4] As explained above, present plans call for expansion of the 1992 eighth grade mathematics TSA exercise pool to include 13 blocks of exercises, including 3 blocks from the 1990 TSA and 10 new blocks.

The exercises in the eighth grade mathematics TSA are organized into five content categories crossed with three process categories. The content areas are (1) numbers and operations; (2) measurement; (3) geometry; (4) data analysis, statistics, and probability; and (5) algebra and functions. The process categories are (1) conceptual understanding; (2) procedural knowledge; and (3) problem solving. The exercises are further classified as calculator-inactive (i.e., exercises for which a calculator would be irrelevant or unhelpful); calculator-neutral (i.e., exercises for which a calculator might be used, but would not be required); and calculator-active (i.e., exercises that students would be quite unlikely to be able to solve without the aid of a calculator). This same framework will be used for both the fourth grade and the eighth grade mathematics assessments in 1992 as well, although of course the proportions of exercises in different content and process areas will differ by grade level (NAEP 1988).

---

[4] Some blocks of exercises may be excluded from the BIB-spiraling. In particular, blocks of exercises that are paced by an audiotape presenting each exercise and telling students when to move to the next one (*tape-paced*, as opposed to *self-paced*) cannot be spiraled because all students in the room must be working on the same exercise at the same time. The three blocks of "estimation" and "higher-order thinking" exercises included in the national NAEP samples but not the TSA samples were excluded from the BIB spiraling for that reason.

The exercise framework for reading in 1992 is organized according to three primary reading purposes and three processes. The purposes are reading for literary experience; reading for information; and reading to perform a task. The process categories are fluency; constructing meaning; and elaborating and responding critically. Three types of materials are used, corresponding closely to the three purposes: literary texts, informational texts, and documents (Mullis et al. 1990).

Because BIB-spiraling results in different students responding to different sets of exercises, item response theory (IRT) models and methods are used to estimate individual-level ability distributions in a common metric, regardless of which booklet a student responded to. Separate scaled score distributions are produced for each content area (mathematics) or purpose (reading).[5] A weighted composite of these separate scales is then created to characterize overall proficiency. Because they are based on larger numbers of exercises and student responses, the estimates for overall proficiency are more accurate than those for subscales. Roughly speaking, both overall and subscale proficiency distributions may be reported at the state level, but subscale scores may not be accurate enough to report for subpopulations within states.

The subscale and overall proficiency score distributions estimated for individuals and groups provide the most useful performance summary for the NAEP cognitive exercises. In addition, however, estimates of difficulties (i.e., proportions of examinees responding correctly, or *p values*) are also routinely calculated and may be useful for special data analytic or reporting purposes. For example, performance on selected "released exercises" is reported to illustrate the meaning of the anchor points established for the NAEP proficiency scales.

---

[5] Note that the NAEP methodology yields a posterior achievement *distribution* for each student on each scale, not a point estimate. Multiple imputed values (the five "plausible values") are sampled from this posterior distribution and used for subsequent analyses. This procedure permits more accurate estimation of standard errors than would the use of a point estimate like the mode or expectation of the posterior distribution.

Sampling

Virtually every substantive inference made from the NAEP data will rest on the assumption that the students actually responding are representative of some larger population. The size and quality of the sample are critical determinants of the imprecision and bias in estimates of population parameters. Imprecision refers to the random error due to the operation of chance in selecting the sample. In principle, it can be made arbitrarily small (i.e., parameter estimates can be made arbitrarily precise) simply by increasing the size of the sample. Bias refers to systematic error due to the sampling design or procedures. For example, some subgroups of students may be underrepresented or overrepresented because of errors in framing the sample, because some sampled schools refuse to participate, or because some sampled students are deliberately excluded for inappropriate reasons or simply cannot be located for testing. In general, bias cannot be reduced simply by increasing the sample size.

As explained above, NAEP includes many samples of students for cross-sectional and bridge studies, at different grade levels and in all participating states, as well as for the nation as a whole. NAEP relies on complex, deeply stratified cluster sampling designs, often with deliberate oversampling of selected subpopulations. In order to estimate population parameters, the responses of participating students are then weighted inversely according to their probability of being selected. It is virtually impossible to draw such samples and derive the sampling weights in a way that yields population estimates entirely free from bias, although the nature and magnitude of any biases may be impossible to quantify. For this reason, when different samples are to be related to one another they should be drawn following the same or highly similar sampling plans. This helps to assure that any unknown biases will be similar across samples, permitting more accurate comparisons of performance from one sample to another.

This section describes some differences between NAEP's national cross-sectional samples and the TSA state samples, as well as potential disparities among the different state samples. Unfortunately, it is generally impossible to quantify the biases these differences and disparities may have given rise to. The NAEP sampling designs are state-of-the-art, and biases due to sampling are probably small. Nonetheless, they pose a potential threat to the validity of inferences from the TSA data. This discussion is framed in terms of the state samples for the 1990 TSA in eighth grade mathematics and the corresponding national sample. Issues for the 1992 TSAs will be much the same.

There are some explicit differences between the national cross-sectional sample and the state samples. National NAEP included both public and private schools and in fact oversampled private schools to provide good estimates of performance by sector. The state samples were limited to public schools. The national sample was augmented with 9-, 13-, and 17-year-old students not in the 4th, 8th, or 12th grades respectively, so that performance could be estimated for both grade levels and age cycles.[6] The TSA samples were for grade levels only (Beaton 1989). Clearly, it is straightforward to adjust for these differences by comparing state data to the grade-eligible public school subsample of national NAEP.

It is more difficult to adjust for some less obvious differences between the state and national samples. First, selection of the national sample began with the drawing of counties or groups of contiguous counties. All public and private schools within each of these primary sampling units were then listed, and schools were sampled from those lists. This procedure was necessitated largely by the difficulty of locating complete rosters of private schools. For the TSA samples, each participating state provided a roster of all of its public schools, and participants were selected directly from those lists.

Second, state samples were designed simply to be representative of each state as a whole, whereas national samples were designed to permit accurate estimation for various demographic subgroups. These different purposes, as well as differences in the demographics of various states and the different information sources used to frame the state (public school) samples versus the national (public and private school) samples, implied that different stratification variables and sampling probabilities were used across states as well as between the states and the nation. In order to assure the inclusion of sufficient numbers of Hispanic students nationally, for example, there was some oversampling of areas where the proportion of Hispanic students was higher than average.

---

[6] In 1990, the national eighth grade mathematics sample was also augmented to include sufficient numbers of examinees from each state participating in the TSA so that an adequate state aggregate comparison (SAC) sample could be constructed, as explained below.

413

This is an entirely acceptable procedure, as good as any practical sampling approach known, but it may have introduced some biases simply because Hispanics in predominately Hispanic areas may not respond to cognitive and background questions in quite the same way as Hispanics in predominantly black or white areas. The fact that such biases are not quite the same from one sample to another inevitably limits the accuracy with which small effects or small performance differences can be estimated.

Data Collection

There were some potentially significant differences between the data collection procedures followed for the national NAEP versus the TSA in 1990, and similar differences will occur in 1992. The evidence available to date suggests that the effects of these differences were generally small. Moreover, from the outset plans were made to scale the TSA and national data separately and adjust for any differences due to procedure using test equating methods, as described below in the section on "Analysis, Reporting, Scales, and Standards." It must be understood, however, that at best, such adjustments can only eliminate the average effect of different administration procedures. To the extent that procedural effects varied from one state to another, the accuracy of both state-to-nation and state-to-state comparisons may be diminished.

The national NAEP has been administered by personnel hired and trained by either ETS or Westat (a subcontractor to ETS primarily responsible for sampling and data collection), or both. For the 1990 assessment, the national data collection was carried out by 35 field supervisors and 90 exercise administrators working over a period of about four and a half months. For the TSA data collection, about 40 state coordinators were responsible for securing cooperation from participating schools, and about 4,000 local administrators (LAs) each participated in a one-day training session and then managed the data collection in a single site (Goodison 1989).

In addition, the scheduling of the national and TSA data collections differed. This is important because additional months of instruction might be expected to confer some small advantage in responding to NAEP exercises, especially in content areas like mathematics, where new topics and procedures may be introduced throughout the year, and at lower grade levels. In 1990, the national data collection lasted from January through May, and the TSA data collection occurred during four weeks in February. To the extent that the effects of extra weeks of instruction are uniform across states, this differences will be compensated for in the TSA-to-SAC equating, described below. If a specific mathematical topic were typically taught before February in some states and after February in other states, however, differential timing effects might be created for exercises testing that topic. Such effects would not be eliminated by the equating procedure.

The preliminary evaluation report on the TSA suggests that overall, cooperation and compliance with instructions were quite good, but that there may have been some isolated problems. Given the small number of states, and the even smaller number that participated in the TSA, however, even a few unreliable data points seriously jeopardize substantive conclusions from state-level analyses. Moreover, publication and discussion of the 1990 TSA results could conceivably lead in 1992 and thereafter to more students receiving "pep talks" or material incentives to participate, or to other irregularities. These are simply manifestations of one of the most fundamental, most intractable differences between the national and TSA data collections: the expectations, purposes, stakes, and motivational contexts of these two data collections differ significantly. High scores matter a great deal more in the TSAs than in national NAEP, and as the TSA results are reported and discussed, they may take on even greater significance in the future.

The NAEP Exercise Pools

As stated above, the heart of both the national assessments and the TSAs is the pool of cognitive exercises administered to students. Inferences from NAEP data are fundamentally limited by the quality and scope of the questions students are asked to answer. By statute, the NAEP exercise pools must be developed following a consensual process. Objectives frameworks for both the 1990 and 1992 TSAs were developed under the auspices of the CCSSO. Exercises conforming to these frameworks were then developed (or will be developed) by ETS, subject to further review and approval by the panels convened by the CCSSO (Mullis 1989). Despite the effort and care devoted to the exercise development process, there are serious constraints on the quality and scope of the final exercise pools. In this section, the 1990 eighth grade mathematics exercise pool is discussed first, then a few comments are offered on the 1992 exercise pools in both reading and mathematics.

First of all, the design of the NAEP data collection requires that all exercises be fairly brief and individually administered, with only limited reliance on nonprint materials (e.g., calculators). Exercises requiring hand scoring are included, at considerable cost, but even these exercises must be designed so that their scoring is as efficient as possible. It is simply not considered feasible to use an exercise that requires five minutes of a scorer's time to determine, say, the solution approach or approaches a student employed.

Second, by and large, the exercise pool must be centered on the proficiencies of the examinee population as they are today. There is a general perception that American students' mathematics skills are inadequate and that some important areas, especially complex reasoning, novel problem solving, and other aspects of "higher-order thinking" are largely neglected. It is simply not efficient, however, to include more than a very few exercises that the vast majority of students will be unable to answer correctly. The TSA consensus planning process may have heightened this difficulty. Despite the rhetoric of avoiding an exercise pool representing the "lowest common denominator," representatives of different states were, understandably, sometimes reluctant to see exercises included that they did not believe the students in their states would be able to answer.

The limitations of the NAEP exercise pool in mathematics are highlighted by a comparison of the NAEP objectives (NAEP, 1988) with the National Council of Teachers of Mathematics (NCTM) *Curriculum and Evaluation Standards for School Mathematics* (NCTM 1989). The NCTM document, which reflects a broad and strong consensus of the mathematics education community, certainly had some influence on the development of the NAEP exercise pool, but the NCTM *Standards* were too forward-looking to be incorporated in their entirety. Moreover, some areas in the *Standards* would be difficult to assess within the constraints of the NAEP design.

The NCTM *Standards* for grades 5-8 set forth five overall curricular goals: Students should learn to value mathematics, become confident in their abilities, become mathematical problem solvers, learn to communicate mathematically, and learn to reason mathematically. Increased attention should be given to open-ended problems; to discussing, reading, and writing about mathematical ideas; and to connecting mathematics to other subject areas and applying it in different contexts, among other emphases. In the light of this vision, the NAEP exercise pool appears quite limited. Although some so-called "higher-order thinking" exercises are included, they fall short of the kinds of open-ended problems envisioned in the NCTM *Standards*, and areas like communication about mathematical ideas and connections of mathematics to other areas are scarcely represented at all. Most of the exercises in the 1990 mathematics pool are multiple choice, and even among the constructed response exercises, only an extremely small number require expository responses as opposed to mathematical notation.

The 1990 TSA exercise pool was even more seriously limited than the national NAEP exercise pool in eighth grade mathematics. For the national data collection, 10 blocks of cognitive exercises were used, and of these only 7 were included in the TSA exercise pool. The 3 omitted blocks were largely concerned with "higher-order thinking" and with estimation. In summary, the 1990 TSA in mathematics can at best indicate students' absolute and relative proficiency with respect to fairly low-level, routine varieties of mathematics problems. It follows that the TSA will probably be insensitive to any effects of mathematics reforms that move curriculum and instruction away from present conventional approaches. If increased emphasis on more complex learning outcomes led to decreased time spent in routine problem solving, it is not unlikely that a state's adoption of a forward-looking mathematics curriculum framework would lead to decreased performance on the TSA exercises. Perhaps most important, the limitations of the NAEP exercise pool will largely preclude the differentiation of distinct learning outcomes. It may be informative to examine results by content area or process category or to compare performance patterns for calculator-active, neutral, and inactive items, but the NAEP mathematics exercises will not enable inferences about progress toward the bold vision of the NCTM *Standards*.

Similar concerns are likely to limit the inferences that can be made from the fourth grade mathematics and reading assessments in 1992. Because the field of reading presently lacks a document setting forth a consensus on curriculum comparable to the NCTM *Standards*, it is more difficult to highlight specific, important reading topics that may be neglected. It is likely, though, that many reading curriculum specialists will wish for longer reading passages and for more questions calling for extended written responses. Some may question whether it is meaningful or even possible to assess "reading for literary experience" in the context of a written examination, where the patent purpose of the reading is to answer examination questions. The area of "reading to perform a task," largely concerned with the use and interpretation of charts, tables, and other documents, represents a relatively new reading construct. Because there is less experience in developing test questions of this kind and because interpretation of documents is rarely taught as a coherent curriculum area, it remains to be seen whether exercises in this area will cohere sufficiently to define a meaningful proficiency scale.

Analysis, Reporting, Scales, and Standards

The methods used by ETS in their analysis, scaling, interpreting, and reporting of TSA findings will significantly shape the inferences that can and will be made from the data. Even though public-use data tapes will eventually become available and other investigators will be able to conduct their own analyses, this will not occur for at least several months after the initial release of findings. Moreover, other scholars are likely to rely heavily on statistics developed by ETS, especially the "plausible values" for each examinee for each scale and subscale. Public-use tape users will also be somewhat handicapped by the masking of certain identifying information, required to assure anonymity of participating students and schools, and to discourage prohibited forms of reporting for geographic areas smaller than states. The procedures followed in logging in, keying and verifying, scaling, and reporting the national and TSA NAEP data are enormously complex, and no attempt is made here to summarize them. This section merely draws attention to a few features of these procedures that may have implications for inferences that can or cannot be made from the TSAs.

TSA data will be pooled across states, then scaled using IRT methods, independent of the national data. The procedures to be used are somewhat complicated by the division of each state's TSA sample into two randomly equivalent subsamples, one of which was monitored. Fortunately, preliminary results suggest that the effect of monitoring was small, so it may be possible for analytical purposes to ignore this distinction. Independent of the scaling of the TSA data, a state aggregate comparison (SAC) subsample of the national NAEP data will be prepared, consisting of grade-eligible public school students in all participating states who were tested at roughly the same time as the TSA data collection occurred. A statistical adjustment will then be made to the TSA scale to align the aggregate TSA score distribution[7] with the SAC score distribution (Johnson 1989). The planned scaling and equating should in principle adjust for systematic differences arising from national versus state administration procedures, motivational contexts, and sampling designs. As pointed out earlier, however, such a global equating can only remove overall average effects. To the extent that administration, motivational, or other effects differed from state to state, they will be neither detected nor compensated.

---

[7] For purposes of equating, TSA data from Guam and the Virgin Islands will not included. The TSA aggregate sample will include only data from the 37 actual states and the District of Columbia.

The NAEP proficiency scales will continue to be the major vehicles for reporting NAEP achievement at both the state and national levels. These scales, derived using IRT procedures, express the achievement of all examinees in a common metric, regardless of which particular exercises they responded to. Scores on each scale range from roughly 100 to 400. As an aid to interpretation, "anchor points" have been established along each scale. Each anchor point is characterized by a label, a brief narrative description of the skill level it represents, and one or more illustrative exercises, which an examinee at the ability level represented by the anchor point would be expected to answer correctly at least 65 percent of the time and which an examinee at the next lower anchor point would be substantially less likely to answer correctly.

Strictly speaking, the IRT scaling methods used to create the proficiency scales depend on an assumption that all the exercises included measure exactly the same ability (are "unidimensional"). The NAEP exercises clearly violate this assumption, and so the IRT-based proficiency scales are better regarded as a useful tool for reducing and summarizing a large amount of data than as the product of a rigorously justified statistical model. Robert J. Mislevy has offered the analogy that a proficiency scale represents a sort of "market basket" of skills, analogous to the market basket of goods used in calculating the Consumer Price Index (CPI). Just as the costs of particular goods and services may vary among cities or change at different rates, so patterns of attainment may differ among groups or over time for specific skills represented in the NAEP exercise pools. The CPI, like a NAEP scale score, gives a global summary, the meaning of which depends on the mix of specific items included. Heavy reliance on the NAEP proficiency scales for TSA score interpretation appears inevitable. These scales invite rankings and comparisons, but taken alone cannot illuminate patterns of strengths and weaknesses across the various subdomains represented by the exercise pool.

TSA results may be reported in various ways. At a minimum, state mean proficiency levels may be reported for each content area and for the overall scale, and proportions of students at or above each anchor point may be reported. Reporting may be enriched by providing co-statistics, as discussed above.

It is also likely that state performance will be reported relative to the three NAGB performance standards for each content area and grade level. The "proficient" level was intended to represent "solid academic performance for each grade tested--4, 8, and 12--and reflect a consensus that students reaching such a level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling." The "advanced" level was to denote "superior performance beyond proficient grade-level mastery," and the lowest, "basic" level was to denote "partial mastery of the knowledge and skills that are fundamental for proficient work at each grade" (NAGB 1990, 5). Efforts to establish these levels for eighth grade mathematics are ongoing, but some significant concerns have been expressed about both the standard-setting process and the meaningfulness and appropriateness of the resulting levels. If state results are in fact reported in terms of the NAGB standards, the reporting will probably be in terms of the proportions of students in each participating state who are at or above each of the three levels. It is clear that the strongly normative descriptions of these levels invite inferences about the absolute level of performance in each state, suggesting that performance below "proficient" is inadequate. It is also clear that the validity of such inferences depends critically on the validity of the NAGB standards.

Warranted and Unwarranted Substantive Inferences from the TSAs

This section takes up some types of inferences that may be attempted using the TSA data and considers their validity in the light of the preceding background. Each of these inferences begins with a *statistical* generalization from the sample tested to the population it represents, but also involves a *substantive* part, which cannot be strictly justified on statistical grounds. The substantive part of an inference may be quite limited (e.g., that TSA performance indicates the degree to which students at a certain grade level have attained the specific learning outcomes the exercises were designed to measure) or quite bold (e.g., that TSA performance indicates the impact of a particular reform initiative on achievement across content areas and grade levels).

Both parts of an inference must be scrutinized. As indicated in the section above, the problems of determining the relation of sample statistics to population parameters may be highly complex. However, the principles are well worked out, and the degree of confidence one may place in statistical generalizations can be quantified. Examining and evaluating the substantive part of an inference is at least as important, but often less straightforward.

It may be helpful to distinguish two major dimensions often present in the substantive component of inferences: *causal attribution* and *outcome generalization*. Causal attribution is integral to our ways of making sense of the world, and is not automatically invalid or misguided. There is a risk, however, that in considering only one or a few explanations for some pattern of findings, plausible alternative explanations may be ignored. There is also a risk, when complex statistical models are used (e.g., path analytic, multistage regression, or "errors in variables" models), that causal hypotheses implicit in the the statistical models themselves may go unnoticed. Outcome generalization is also inherent in virtually any interpretation of test data. A score is nearly always taken to signify more than mere performance on the particular set of exercises administered. There may be a tendency, though, to take the narrow outcomes assessed in the TSA and generalize too broadly, across grade levels, across learning outcomes or levels of cognitive complexity within a content area, or even to other content areas.[8]

Perhaps the most important, valid, and useful class of inferences that may be made from the TSAs is one for which the substantive part of the inference is minimal. Historically, NAEP has been conceived primarily as a system for documenting achievement levels and trends over time, communicating the meaning of achievement levels largely by referring to the (released) exercises themselves and directing attention to patterns of improvement or decline. This is a worthy purpose, and there is, potentially, considerable value in extending this function in itself to the state level. Given that responsibility for providing public education is vested in the states, the influence of achievement information reported at the state level may be substantially greater than the impact NAEP has had historically. If TSA findings simply characterize states' overall achievement levels and distributions or document achievement disparities among their student subpopulations, implications may follow concerning allocations of resources or research priorities, without any need to attempt to use the TSAs to discover the reasons behind achievement patterns or levels. These implications follow from information about achievement levels in themselves--not from judgments about the quality of the educational system nor the success of policies or programs or reforms, nor even the relative contributions of a state's demographics versus the quality of its educational system as determinants of measured performance. It may be enough to say, "This group of students within the state is performing far better than that group."

---

[8] Causal attributions and outcome generalizations by no means exhaust the range of substantive assumptions that may require scrutiny when considering inferences from the TSAs. As examples, the implications derived from an analysis of some policy may rest on the assumption that its effectiveness is independent of the state context in which it is implemented or that it would have the same effect in the future as it had five years earlier.

The 1990 and 1992 TSAs are very limited in scope. Despite some promising innovations, the mathematics and reading assessments are largely limited to lower-level abilities assessed using mostly conventional exercise formats, at one or two grade levels. They constitute just three or four "narrow beams" with which to illuminate the broad terrain of learning outcomes across grades, curriculum areas, and levels of cognitive complexity. Within the decade, however, if the NAEP exercise pools continue to evolve and if the TSAs are extended to more grades and subject areas, it may become possible to monitor states' achievement levels over time and to relate entire distributions of performance levels within a given state to those in other states, in the nation as a whole, and perhaps even in other nations. Moreover, the interpretive context for state assessment data may include normative standards evolving from the NAGB standard-setting process, providing "absolute" anchors against which to judge students' educational attainments.

There is an increasing demand for data permitting comparisons across states, and it is nearly certain that as such data become available, they will be over-interpreted. Consequently, the discussion to follow in this section is primarily cautionary in tone. Nonetheless, nothing said below should be taken to detract from the value of descriptive information of the kind NAEP has historically provided or from the potential value of making such information available at the state level.

The remainder of this section addresses five general types of inferences from the TSA data. These are inferences about: (1) achievement in the nation as a whole; (2) the overall quality of state educational systems; (3) educational resource allocations; (4) educational reforms, policies, or practices; and (5) more complex, specific questions concerning educational theory, policy, or practice.

Inferences about Achievement in the Nation as a Whole

The TSA data will *not* permit inferences about national achievement levels or patterns. Even though the TSA samples, taken together, far exceed the size of the national NAEP sample, the aggregation of the 37 state samples in 1990 cannot be taken to represent the nation. In theory, mean performance levels in the 37 participating states could all be above the national average, although that is unlikely.[9] It is an open question whether eventually, if all 50 states participate, the national NAEP sample might be supplanted by the collection of state samples, but such a decision should not be taken lightly. Initially, state NAEP results are likely to improve for a number of years simply as a result of direct pressure for higher scores (Haertel 1989; Haney and Madaus 1986). Reliance on state NAEP in lieu of a separate national assessment could jeopardize fragile trend lines that have been maintained in some content areas for over 20 years.

Inferences about the Overall Quality of State Educational Systems

The question of which states have the best educational systems is likely to be among the first asked of the TSA results, but attempts to answer it are fraught with perils. State-to-state or state-to-nation comparisons with respect to overall educational quality will involve generalizations from the TSA exercise pools to a broader range of learning outcomes and from the grade levels assessed to other grade levels, as well as attributions of TSA results to educational quality as opposed to out-of-school influences. Judgments about the absolute levels of achievement in a single state will depend critically on the adequacy of normative performance levels. As discussed above, the 1990 TSA data, for example, can only inform a limited range of mathematics learning outcomes, at one grade level. Patterns of performance across states might differ dramatically for other content areas or grade levels, or even for unmeasured learning outcomes in fourth or eighth grade mathematics.

Even if states can be ranked with respect to the specific content of the NAEP scales, the relevance of those rankings to judgments about educational quality will depend on the match between the NAEP content and the state's own curriculum framework. Differential test-curriculum match across states poses an additional threat to the validity of inferences about overall educational quality. As already noted, for example, a state pressing to implement the NCTM framework might experience a (possibly temporary) decrease in performance on conventional math problem-solving exercises, due to its deliberate decision to allocate less instructional time to that type of problem. The 1990 TSA might support the (valid) inference that the state's performance on that type of problem was lagging, but not the (invalid) inference that their overall mathematics performance was lagging. A state's determination that the TSA content was relatively less important than some learning outcomes not assessed, or that the teaching of algorithmic skills should be deferred to a later grade level, certainly ought to condition some interpretations of its assessment results.

---

[9] The aggregate TSA-to-SAC equating will assure that the (properly weighted) average of the TSA means (excluding Guam and the Virgin Islands) is equal to the (properly weighted) average performance of public-school grade-eligible students tested at about the same time as part of national NAEP, in the same 37 states and the District of Columbia.

In addition to questions of the generalizability and relevance of the content assessed in the TSAs, inferences about educational quality involve, at least implicitly, attributions of performance differences to causes within the purview of the educational system. Plans for reporting and interpreting TSA data call for the use of co-statistics to rank states according to the educative difficulty they face or otherwise adjust for demographic differences, but the long history of educational policy research attempting causal inferences from nonexperimental data attests to the problems inherent in such adjustments. Moreover, the problem is exacerbated in the TSA data by the small number of data points available and the possibility that some of these data points are outliers.

In future years, if NAEP can provide state-level trend lines, stronger inferences may be warranted. As already noted, however, immigration from abroad and migration from state to state could produce sharp, unpredictable fluctuations in achievement levels that had nothing to do with school policies or practices. Changes in financing and incentive structures may also induce migrations between public and private schools, further complicating the interpretation of a state's public school achievement trend lines. Even apart from these difficulties and even given time-series data, the complex environment in which educational policy decisions are made and implemented makes causal inference highly problematical.

Thus far, this discussion of inferences about the overall quality of state educational systems has addressed "relative" comparisons among states or between states and the nation as a whole. An alternative, "absolute" characterization is also possible. If the NAGB performance standards are adopted, participating states' performance will probably be characterized in terms of the proportions of examinees at or above each of the basic, proficient, and advanced levels. If only a small proportion of a state's students are estimated to fall in the higher performance categories, the state's educational system may be judged inadequate, regardless of the limitations of the exercise pool, the out-of-school backgrounds of the state's student population, or any controversy surrounding the performance standards themselves.

Inferences about Educational Resource Allocations

If the TSA data can inform judgments about the relative success of the educational system in meeting two or more educational goals, they may have implications for the allocation of educational resources among those goals.[10] The set of goals might involve different learning outcomes within or across content areas, or might involve the same learning outcome for two or more groups of students. Under the present TSA reporting constraints, student groups might be defined by race or ethnicity, for example. If prohibitions on reporting below the state level were lifted, groups might also be defined by, say, district boundaries. Clearly, allocation decisions have a normative component, and cannot be reduced to purely technical problems. Achievement data can do no more than inform policy makers' opinions.

In principle, this use of TSA data would not require any specific causal attribution, and might involve only limited generalization beyond the content actually assessed. It would, however, require that at least two distinct learning outcomes be measured. As discussed above, the quality of the NAEP exercise pool is such that many important content distinctions are difficult to support, such as that in mathematics between solving routine, algorithmic problems and complex, open-ended problems.

There may be some risk that if the TSAs are expanded to include more content areas and if normative standards are established for each of these areas, then relative proportions of students meeting high standards in different areas will be taken to indicate the relative degrees of investment required in those areas. Inferences of this kind would, of course, depend critically on the defensibility and comparability of the standards, within and across content areas. A higher proportion of "advanced" fourth graders in reading than in mathematics would only be meaningful to the extent that the cutting scores taken to define "advanced" performance in the two content areas were grounded in some common conception of the meaning of "advanced" performance across content areas.

---

[10] A judgment based on overall achievement levels that increased resources should be devoted to education versus other societal needs is also an allocation decision, but of a kind not addressed in this section.

The use of the TSA data to influence resource allocation across student groups appears more promising and more salutary. As noted earlier, by directing attention to substantial achievement disparities, the TSA might encourage increased attention to those student groups traditionally underserved by our educational system. Again, of course, patterns of measured achievement in themselves carry no necessary implications for action. Policymakers could choose to reward high-achieving schools or districts, to the relative detriment of those already at a disadvantage.

Inferences about Educational Reforms, Policies, or Practices

Another major class of inferences from the TSA data will consider the effectiveness of specific educational policies or instructional practices, or appraise the effectiveness of specific educational reforms. In general, such inferences will be based on analyses comparing groups of states with and without the given reform or policy, or analyses modeling achievement across states as a function of the given policy or practice together with control variables. These analytical approaches are referred to above as "natural experiment" designs. As stated earlier, the fact that the 1990 TSA includes only 37 states and the 1992 data will include 41 states sharply limits the potential of these designs, especially given the enormous heterogeneity among states in both their demographics and their school systems. State-specific problems including low participation or high exclusion rates also threaten the validity of inferences from such designs.

The major substantive (as opposed to statistical) problem in trying to assess specific reform initiatives is the multiplicity of plausible rival hypotheses that might explain statistically significant achievement differences across groups of states. The actual effects of particular policies or practices are likely to be subtle and easily swamped by the enormous variation in outcomes attributable to student demographics as well as other policies and practices. Statistical adjustments for such demographic differences will be indispensable, but as stated above, such adjustments are highly imperfect. Even with adjustments, it may be impossible to distinguish small but robust effects from statistical artifacts (Koretz, in press). It would be most unfortunate if "no effect" findings from such insensitive analyses doomed promising or successful educational practices or programs.

In the future, state-level trend lines might permit "interrupted time series" analyses, seeking, for example, evidence of global effects from major, multifaceted state education reform initiatives. State NAEP data probably could not be used to determine the particular policies or practices responsible for any changes detected and would surely not be sufficient in themselves to evaluate such comprehensive reforms, but they might provide highly valuable baseline and trend data. As with analyses of effects of specific reforms, "no effects" findings from interrupted time-series analyses would very likely be inconclusive.

Inferences about educational policies and practices are also rendered problematical by the narrow focus of the TSA outcome variables. The most promising analyses would be those examining the effects of reforms, policies, or practices closely tied to the specific outcome measures available. For example, eighth grade mathematics outcomes might be related to state adoption of a revised curriculum framework for middle school mathematics between one and four years prior to the assessment. Such analyses will be somewhat opportunistic. Neither NAEP nor the TSAs were conceived or designed with prominent attention to this kind of application. Even such focused analyses might be better viewed as sources of promising hypotheses for further investigation than as confirmatory.

An alternative analytical strategy for addressing questions of this kind would employ less highly aggregated TSA data. If background information, including questions in the SCPQ and teacher questionnaires as well as student background questions, could be related to outcomes in a meaningful way, some useful findings might be obtained. Obviously, this approach could only be used to investigate those policies, reforms, or practices that could be operationalized in terms of responses to background questions.

In principle, of course, the same kinds of analyses could be carried out using national NAEP data, although the history of such analyses using national NAEP data is not especially encouraging (Stedman and Smith 1983). Over the years, NAEP has been most useful for documenting achievement levels and trends over time, and considerably less useful for evaluating policies or practices. Such analyses might be more successful using the larger TSA database, however, if only because a larger number of "control" variables could be incorporated.[11] If the TSAs evolve to the point of providing comprehensive, stable, state-level trend data, they may become quite useful in suggesting hypotheses for investigation in more carefully designed and focused studies. TSA data from different states might be analyzed as independent replicates, lending additional weight to findings consistent across states.

---

[11] Although costatistics defined at the state level could be included among control variables, the most useful control variables would be those defined at the same level of aggregation as the dependent variables. Due to confidentiality restrictions, this effectively limits the available control

Specific Questions Concerning Educational Theory, Policy, or Practice

The TSAs may also be of occasional value in addressing more complex, focused questions. As an instance, if some states but not others had implemented a reform to improve minority mathematics performance at the middle school level, it might be evaluated by comparing 1990-to-1992 changes in the minority-majority achievement gaps for states with versus without that policy, controlling for other state characteristics, including demographic changes from 1990 to 1992. Such analyses will necessarily be quite focused and to some extent opportunistic. To the extent that the TSAs happen to include the outcome and demographic variables needed to inform some current substantive question, they may be quite helpful. Careful attention to the background questions included in future TSAs may enhance the utility of the data for analyses of this kind.

As state-level trend lines are established, the TSA data should become increasingly valuable. A state's own achievement level at some earlier point in time can serve as a benchmark that takes its demographics and even the prior history of its own educational system into account, strengthening the plausibility of attributions for improvement or disimprovement to some constellation of current educational policies and practices, together with possible changes in the characteristics of successive examinee cohorts. The two data points for eighth grade mathematics provided by the 1990 and 1992 TSAs for those states participating in both trial assessments will provide some preliminary indication of how valuable such trend lines may be and how they are likely to be construed. Elected officials may claim credit for any progress shown and may interpret TSA gains as indicating improvement on a broader range of outcomes than was actually assessed. In the light of these possible interpretations, it bears repeating that as the TSA scores take on increased importance, there is likely to be an initial period of some years during which performance improves in part for reasons unrelated to the actual quality of student learning outcomes. If the content assessed in the TSAs is too narrow, an increased curricular focus

---

variables to items included among TSA background questions.

on TSA outcomes might even lead to poorer learning with respect to (possibly more important) outcomes not assessed.

Summary

The 1990 and 1992 Trial State Assessments may initiate the most significant changes in NAEP since its inception over 20 years ago. Enormous effort has been expended over a remarkably short period of time to fashion their content and procedures; create an extensive field organization; administer and monitor the data collection; and now to process, report, and interpret the data collected. It may at first seem surprising, in the light of all this effort, that the purposes the TSAs are to serve are less than entirely clear. This paper has briefly reviewed some elements of the educational policy context that led to the TSAs, the structure and technical characteristics of the data collected, and some of the kinds of inferences, valid and invalid, that may be attempted.

There are a few potentially important technical limitations of the TSA data, but by and large, the TSA is close to the state of the art in large-scale assessments. Even the lowest school-level and student-level participation rates and even the highest student-exclusion rates are really not all that bad. Comparability across states may be somewhat threatened and state rankings may be somewhat influenced by variations across states in the degree of cooperation obtained, but it is very unlikely that sampling or other technical problems will lead to any gross distortions of estimated achievement levels.

The limited content of the TSA exercise pools is a more serious concern. The consensus processes leading to the assessment frameworks were implemented and executed conscientiously and exercise development was highly competent. Nonetheless, the tension between assessing achievement as it is versus as it ought to be, the severe time constraints under which the exercise pools were developed, and both economic and logistical constraints on exercise timing and scoring resulted in an exercise pool for mathematics that some experts have found disappointing. Similar factors seem likely, in the end, to temper the boldness of the 1992 reading framework. The TSAs can directly inform only those learning outcomes actually assessed. The 1990 TSA will provide a limited view of achievement in mostly conventional mathematics at one grade level, in about two thirds of the states. The 1992 TSA will provide about the same kind of information in eighth grade mathematics.

Inferences from the TSA data have both a statistical aspect and a substantive aspect. There is a tendency, perhaps, to dwell more on the problems of valid statistical inference from a sample to a population than on the substantive assumptions required to support interpretations of the meaning of TSA findings or their implications for action. Nonetheless, the substantive part of the inference is by far the weaker. Interpretations of TSA findings may be especially prone to problems of faulty causal attribution and of overgeneralization beyond the content actually assessed.

In my opinion, the most important and most defensible inferences from the TSA data will be those for which the substantive assumptions are weakest, simply documenting achievement distributions at the state level and using selected exercises from the NAEP exercise pool itself to illustrate the meaning of selected score levels. This is no more than an extension of the historical role of NAEP to the state level. The states are responsible for public education, and for that reason, such information may have greater impact when reported at the state level than for the nation as a whole.

It will be especially important to document achievement disparities among subpopulations within states, although the TSA sample sizes may be too small to estimate achievement accurately even for blacks and for Hispanics within some states, let alone smaller subgroups. Achievement gaps between white and minority students may be taken to imply inequities in our provision of a sound education to all of our nation's children, although others may favor explanations that place greater weight on nonschool factors. Along with a range of possible explanations, NAEP findings are likely to be cited as support for a variety of incompatible remedies.

The 1990 and 1992 TSA data will *not* be useful for describing achievement in the nation as a whole, and even if all 50 states eventually participate, the use of the pooled TSA data to supplant the national NAEP sample may be unwise. The TSAs are also likely to be of almost no value in characterizing the overall quality of states' educational systems or the effects of particular policies, reforms, or instructional practices, primarily due to limitations of the exercise pool, limitations on the identification of data below the state level with particular schools or districts,[12] the complex policy environment in which educational policies and reforms are enacted and implemented, and the range of out-of-school factors influencing school achievement, including patterns of migration among states and immigration from abroad. In the future, if NAEP comes to provide state-level trend lines, somewhat stronger inferences may be warranted about the overall effect of major state-level educational reform initiatives.

In principle, two types of allocation decisions might be informed by the TSA data: allocation of instructional time and other resources to one versus another learning objective and allocation of instructional resources to one versus another area or subpopulation. Due to the limited potential of the TSA exercise pools to differentiate among learning outcomes, the data will be of limited usefulness for the first type of allocation decision, but as already discussed, they may have implications for the second type of allocation decision, especially if teacher, school, and student background questionnaires can be used to document differences in the characteristics of teachers serving different student subpopulations, or in the character of the instruction they receive. Finally, the TSA data may be of occasional value in addressing more focused questions relating to the specific areas assessed. Such analyses will be somewhat opportunistic and will depend heavily on the scope and quality of the TSA background data. The usefulness of the TSA data for such purposes will increase significantly over the years as trend lines are established.

Recommendations

- The NAEP exercise pool should be expanded to permit more differentiated reporting of learning outcomes within content areas. In particular, the number of exercises measuring "higher-order thinking" and the number of exercises requiring constructed responses should be expanded. In mathematics, more exercises should require the use of language in responding, and the range of learning outcomes discussed in the NCTM *Standards* should be more fully represented.

---

[12] This and other occasional comments on limitations arising from the prohibition against reporting TSA results below the state level *should not be construed* as advocating such reporting. The question of reporting below the state level is quite complex, and entirely beyond the scope of this paper.

- Careful attention should be given to the reporting of state-level data for demographic subgroups, including minimum sample sizes and levels of statistical precision required to report separate findings for a given group within a state.

- If the TSAs are continued beyond 1992, consideration should be given to state-level sampling plans permitting more accurate reporting for subpopulations, especially major racial-ethnic groups.

- If the TSAs are continued beyond 1992, every effort should be made to provide stable trend data at the state level and, where possible, for within-state subpopulations.

- Background questions and co-statistics should be selected and reported to draw attention to as well as control for plausible rival explanations for patterns of NAEP findings, including private school enrollment patterns, demographic trends, and other factors that might affect the comparability of the populations sampled in different states.

- Strong, explicit caveats should be promulgated concerning the range of plausible rival explanations and interpretations for patterns of NAEP findings, in an effort to discourage attributions of achievement patterns to specific state-level policies or practices, or judgments about "overall quality" of different states' educational systems.

- Within the foreseeable future, the TSAs should not supplant national NAEP samples, even if all states elect to participate.

- Consistent with agreements entered into with participating states and as permitted by statute, TSA data should be made available as quickly as possible and in as much detail as possible to responsible users interested in secondary analysis.

References

Alexander, L., and James, H. T. 1987. *The Nation's Report Card: Improving the Assessment of Student Achievement* (Report of the Study Group). Washington, DC: National Academy of Education.

Beaton, A. E. March 1989. "Sampling Design for the 1990 Trial State Assessment Program." Paper presented at the meeting of the American Educational Research Association, San Francisco.

Goodison, J. March 1989. "The 1990 NAEP Trial State Assessment Data Collection." Paper presented at the meeting of the American Educational Research Association, San Francisco.

Haertel, E. H. 1989. "Student Achievement Tests as Tools of Educational Policy: Practices and Consequences." In *Test Policy and Test Performance: Education, Language, and Culture* Ed. B. R. Gifford. Boston: Kluwer Academic Publishers.

Haney, W., and Madaus, G. 1986. "Effects of Standardized Testing and the Future of the National Assessment of Educational Progress." Paper prepared for the NAEP Study Group. Boston: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Johnson, E. March 1989. "Issues and Procedures in Analyzing the State Assessment Data." Paper presented at the meeting of the American Educational Research Association, San Francisco.

Koretz, D. M. (in press). "State Comparisons Using NAEP: Large Costs, Disappointing Benefits." *Educational Researcher.*

Mullis, I. V. S. March 1989. "NAEP's 1990 Trial State Assessment Program: New Challenges in Instrument Development." Paper presented at the meeting of the American Educational Research Association, San Francisco.

Mullis, I. V. S., et al. April 1990. *The NAEP Guide: A Description of the Content and Methods of the 1990 and 1992 Assessments* (Report No. 21–TR–01). Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress [NAEP]. 1988. *Mathematics Objectives: 1990 Assessment* (Report No. 21-M-10). Princeton, NJ: Educational Testing Service.

National Assessment Governing Board [NAGB]. 1990. *Setting Appropriate Achievement Levels for the National Assessment of Educational Progress: Policy Framework and Technical Procedures.* Washington, DC.

National Council of Teachers of Mathematics [NCTM] March 1989. *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA. (Prepared by the Working Groups of the Commission on Standards for School Mathematics, Thomas A. Romberg, Chair)

Stedman, L. C., and Smith, M. S. 1983. "Recent Reform Proposals for American Education." *Contemporary Education Review*, 2: 85-104.